EDUARDO JOSÉ MARQUES PEREIRA

Humans in Action at Different Levels:
the group, the whole, and the parts

Ph.D. Thesis

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
January 2016

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Humans in Action at Different Levels:
the group, the whole, and the parts

Eduardo José Marques Pereira

Programa Doutoral em Digital Media UT-Austin | Portugal

Supervisor: Jaime dos Santos Cardoso (PhD)

Co-supervisor: Ricardo Morla (PhD)

January 2016

Resumo

Os humanos nunca se encontram sozinhos, eles são observados quer por outros humanos, quer por câmaras em diversos cenários e contextos. Uma quantidade substancial de câmaras de video-vigilância estão a ser instaladas em espaços públicos no decorrer dos últimos anos, devido, em larga escala, a questões de segurança, mas também, e com tendência crescente, para extrair estatísticas comportamentais de pessoas em diferentes locais tais como infraestruturas de transportes públicos, centros comerciais, e áreas de desporto, entre outros. O denominado *visual analytics* tem vindo a ganhar notoriedade na indústria, onde a informação comportamental é valiosa e extremamente desejada pelos clientes nas áreas de retalho, desporto e entretenimento, só para citar algumas.

A enorme quantidade de informação visual não está só a ser gerada pelos vídeos de video-vigilância. Especialmente, aparelhos portáteis, como os telemóveis, estão a facilitar a produção massiva de conteúdo digital que está constantemente a ser armazenado na internet através do uso de redes sociais digitais como o Facebook. Tão larga coleção de vídeos precisa de soluções automáticas que eficientemente retorne informação a partir de etiquetas pré-definidas, podendo ser filmagens de determinados tipo de ação ou atividade.

A compreensão do contexto do ambiente visual é essencial para adequadamente interpretar o comportamento, uma vez que o contexto é diferente para cada tipo de aplicação. Filmes de ficção ciêntífica usualmente contêm visões de espaços ubíquos que compreendem o comportamento e antecipam a necessidades dos habitantes desses espaço. A análise visual eficaz de símbolos não verbais e de expressividade é central para uma correta interação e comunicação nesses meios inteligentes.

Analisando com atenção os domínios aplicacionais supracitados, *video-vigilância*, *multimedia* e *comportamental*, a análise de atividade humana é fulcral em todas elas. No entanto, cada uma providencia diferentes restrições e contexto. Esta dissertação apresenta um olhar abrangente sobre o ecosistema da análise de atividade humana na área da visão computacional, sugerindo uma categorização de ações em três níveis diferentes, *o grupo* na cena, *o todo* na imagem, e *as partes* no corpo, definida pelo domínio em que a aplicação reside. Desta forma, procuramos por uma caracterização intermédia que forneça uma ponte natural entre o tipo de conteúdo da aplicação, i.e. as entradas perceptuais, e as necessidades da aplicação, i.e. as inferências. Assim sendo, esta tese investiga representações de movimento e de contexto relacional, como sendo as entradas perceptuais, para suportar a modelação e deteção de ações humanas a cada nível pré-definido, como sendo as inferências.

Esta tese segue uma metodologia orientada à aplicação que define os requerimentos da aplicação a o processo de avaliação. Para cada aplicação é proposta uma *framework* que engloba uma implementação completa e funcional, e que também integra soluções para perguntas de investigação previamente identificadas e definidas pelo domínio aplicacional. Várias experiências são relatadas e servem de suporte de evidência para as contribuições funcionais e académicas.

Abstract

Humans are never alone, they are being observed either by other humans or by cameras in many different scenarios and contexts. Substantial amount of surveillance cameras have been deployed in public spaces in recent years, largely in response to security and safety concerns, but also, and gradually evolving, to extract behavioral statistics of people in different venues such as transport infrastructures, shopping centres, and sport arenas, among others. The so-called *visual analytics* is broadly gaining interest in the industry, where valuable behavioral information is extremely desired by clients in the retail, sport and entertainment areas, just to name a few.

The huge amount of visual information is not only being generated by surveillance videos. Specially hand-held devices, such as smartphones, are facilitating a massive production of digital media content that is constantly uploaded to the internet using multimedia social platforms such as Facebook. Such a large video collection needs automatic solutions to efficiently retrieve information from pre-defined labels that may be footages of a certain type of actions or activities.

Understanding the context of a visual environment is essential to properly interpret behavior, since the context will be distinct for each application. Science fiction movies often contain visions of ubiquitous spaces that understand the behavior and anticipate the need of their inhabitants. Effective visual analysis of non-verbal signs and expressiveness is central to meaningful interaction and communication in such intelligent environments.

Looking carefully to the aforementioned domain settings, *surveillance*, *multimedia* and *behavioral*, visual human activity analysis is crucial for all of them. However, each one provides different constraints and context. This dissertation provides a bird-eye of the ecosystem of human activity analysis in computer vision by suggesting the categorization of actions in three different levels, *the group* in the scene, *the whole* in the frame, and *the parts* in the body, defined by the domain settings in which the application resides. In this way, we look for an intermediate characterization that provides a natural bridge between the type of content of the application, i.e. the perceptual inputs, and the application needs, i.e. the inferences. Therefore, this thesis investigated motion and relational-context representations, as the perceptual inputs, to support the modelling an detection of human action at each defined level, as the inferences.

This thesis follows an application-based methodology to define the application requirements and evaluation process. For each application is proposed a framework that encompasses a complete and functional implementation, and that also integrates solutions to research questions previously identified and defined by the domain settings. Several experiments are reported to support the evidence of both the functional and academic contributions.

Acknowledgments

A long walk, a long journey ... everything started when I was four years old ... no, come on, I am just joking. But, honestly?! Yes, it was ... a long ... but exciting and challenging stage of my life. A great deal has happened in my life since I engaged on this adventure, therefore human activity analysis for me was not only my research topic. In research, we are always looking for certainty, for the best optimization, for the fastest and most reliable method, however, in life, in the social life, the best part is the uncertainty, the moment of surprise, the perfect moment that can last for a couple of minutes but that will be always present for us.

During this period, I had the chance to talk, meet, interact and work with remarkable people. Thanks to my advisors, Prof. Jaime Cardoso and Prof. Ricardo Morla. My sincere gratitude to Prof. Jaime for being a model to follow in the research field, and for his support and confidence along my thesis. To Prof. Pimenta Alves my admiration for the youngest mind that I have ever known, and also my gratitude for his unconditional support. To my colleagues of the VCMI group, sorry, my mistake ... my friends, especial thanks to Ana, Filipa, Inês, Hélder, Ricardo, Lucian, Pedro, João, Luis, and Kelwin ... and for a special member Filipe. All of you have been extraordinary! Thanks, indeed. To the PhD students of the social-psychology team from the University of Porto and to the staff of the Agrupamento de Escolas Eugénio de Andrade, Escola EB2/3 de Paranhos for their help and advise in our work. Of course, many many thanks to Kelly Rodrigues and Ana Rio. To Sensei Shini'chi Satoh from NII, Japan, for his wise advices and guidance during my extraordinary stay in the wonderful city of Tokyo. To my favourite vietnamese friend Sang, and my colleague from Iran but living in Canada, Mehran. To David, Santiago, Andrés and Prof. Castellanos-Dominguez from the SPRG, Colombia for our collaborative work. To Pedro Ângelo and, specially for his last support, Bruno Oliveira, friends for ever ... Last but not least, to INESC TEC, specially to the members of the board direction of the CTM's department ... and of course FCT for his support.

Finally, to my family, my parents Luisa and Joaquim, this is for you ... without your support this path would be much more difficult to follow. To my sisters, Sophia and Gabriela, because they love me and I love them. To my brother-in-law, Ricardo for his happiness. And, finally, to my nephews, Gustavito and Gabriel ... especially you Gabriel, your role these last months has been very important for me, thanks ;)

I hope not to have forgotten anyone ... Now, the next step, move on ...

Eduardo M. Pereira
Porto, July 26, 2016

"I am seeking.
I am striving.
I am in it with all my heart."

Vincent Van Gogh

Contents

List of Figures

List of Tables

Acronyms

A	Accuracy
ADAS	Advanced Driver Assistance Systems
AE	Accumulated Error
ASC	Autism Spectrum Conditions
ASLLVD	American Sign Language Lexicon Video Dataset
BACK-F	BACKward-Feature selection
BEAST	Body Expressive Action Stimulus Test
B.I.	Balanced Interests
BN	Belief Network
BoF	Bag of Features
BoW	Bag of Words
BSLCP	British Sign Language Corpus Project
CCTV	Closed Circuit TeleVision
CENTRIST	CENsus TRansform hISTogram
CFG	Context-Free Grammars
CHAT	Chatting
CHD	Curve segment Hausdorff Distance
CM	Confusion Matrix
CNN	Convolutional Neural Networks
CPD	Conditional Probability Density
CR	Correct Rate
CS	Crowd Structured
CU	Crowd Unstructured
Cut	Cut Transition
CKA	Centered Kernel Alignment
DBN	Dynamic Belief Network
DCM	Discrete Choice Model
DCS	Divergence-Curl-Shear
DevNet	Deep Event Network
DFT	Discrete Fourier Transform
Dist.	Distracted
Dis.	Disoriented
DPM	Deformable Part Models
DTW	Dynamic Time Warping
E.I.	Equally Interested
EM	Expectation-Maximization
Exp.	Exploring
F_1	F1-score
FABO	Bimodal Face and Body Gesture Database
FAST	Features from Accelerated Segment Test
FCM	Fuzzy C-Means

FN	False Negative
FM dataset	Friends-Meet dataset
FOE	Focus of Expansion
FP	False Positive
FSCG	Free-Standing Conversational Group
FSM	Finite-State Machine
FV	Fisher Vector
GB	Geometric Blur
G.B.	Group Behavior
GDSR	Group Detection Success Rate
GMM	Mixture of Gaussians
GNC	Graduated Non-Convexity
GraT	Gradual Transition
GT	Ground Truth
HBA	Human Behavior Analysis
HCI	Human Computer Interaction
HDP	Hierarchical Dirichlet Process
HMM	Hidden Markov Models
HOF	Histogram of Optical Flow
HOG	Histogram Of Gradients
IID	Independent and Identically Distributed
IIT	Israel Institute of Technology
Int.	Interested
I.P.	Individual Profile
IP camera	Internet Protocol camera
IQR	InterQuartile Range
LCS	Longest Common Subsequence
LDOF	Large Displacement Optical Flow
LDS	Linear Dynamical Systems
LPD	Lagrangian Particle Dynamics
LPT	Lagrangian Particle Trajectories
LSK	Local Steering Kernel
LSA	Latent Semantic Analysis
LSWMS	Line Segment Weighted Mean-Shift
LTA	Linear Trajectory Avoidance
mAP	Mean Average Precision
MBH	Motion Boundary Histogram
MCMC	Markov Chain Monte Carlo
MED	Multimedia Event Detection
MG	Motion Gradient
MHI	Motion History Image
MIL	Multiple Instance Learning
MLD	Moving Light Display
MLP	Multi-Layer Perceptrons
MOTA	Multi-Object Tracking Accuracy
MOTP	Multi-Object Tracking Precision
MRF	Markov Random Field
MSAC	M-estimator SAmpling and Consensus
MTS	MultiTracking Sparse
MTD	MultiTracking Dense

NCC	Nonverbal Communication Computing
NFA	Nondeterministic-Finite-State Automaton
NIST	National Institute of Standards and Technology
NN	Neural Network
ODE	Ordinary Differential Equation
P	Precision
PCA	Principal Component Analysis
PCH	Pixel Change History
PDF	Probability Density Function
PDT	Pedestrian Detection and Tracking
PN	Petri Net
PPHT	Probabilistic Hough Transform
PPN	Probabilistic Petri Net
PSE	Pixel Signal Energy
PSL	Portuguese Sign Language
QoM	Quantity of Motion
R	Recall
RBF	Radial-Basis Function
ROI	Region Of Interest
RKF	Runge-Kutta-Fehlberg
SBD	Shot-Boundary Detection
SFM	Social Force Model
SFV	Stacked Fisher Vector
SIFT	Scale-Invariant Feature Transform
SOM	Self-Organizing Map
sPAcCT	spatial Principal component Analysis of color Census Transform histogram
SPRG	Signal Processing and Recognition Group
SSP	Social Signal Processing
ST-tube	Spatio-Temporal tube
STIP	Spatio-Temporal Interest Points
SVM	Support Vector Machine
SWLDA	StepWise Linear Discriminant Analysis
TB	True Balance
TLD	Track Learn Detect
TP	True Positive
TREC	Text Retrieval Conference
TRECVID	Text Retrieval Conference's Video Retrieval Evaluation
TV	Total Variations
U.I.	Unbalanced Interests
VLAD	Vector of Locally Aggregated Descriptors
VIBEA	VIdeo-based Body Expressiveness Analysis
VILOMA	VIdeo-based LOngr-range Motion Analysis
VIMUAR	VIdeo-based MUltimedia Action Recognition
VISOBI	VIdeo-based SOcial Behavior Identification
VRA	Variance-based Relevance Analysis
KLT	Kanade-Lucas-Tomasi
k-nn	*k*-Nearest Neighbours
KRAV	Kernel-based Relevance Analysis for Video data

Chapter 1

Introduction

> Inspiration does exists but it must find you working.
>
> Pablo Picasso

Humans are everywhere. They exists, they move, they interact with each others, they express themselves in multiple ways and in different social contexts, they are alive. There is hardly a single place on earth where one cannot find evidences of human presence. Wherever they are, they bring their desires and their intentions. And intentions lead to motion. Contrary to the saying, it is not money that makes the world go 'round. Intentions do. Intentions generate motion, and humans are in perceptual motion, staging their intentions. Either solo or in group, humans behave according to their desires, with a more or less established plan. The motion by desire is so engraved in humans that if one looks carefully, with a trained eye, it is even possible to infer the intentions from the behavior, the motion, of a single or a group of humans. Trained professionals are used to infer human behavior in different settings such as suspicious activities in public spaces, violent scene crimes, customers' experience in retail spaces, social action analysis, among others. But humans are humans and are used to see each other behave for millenia. They can, consciously or unconsciously, spot when something is not quite right, or something is wanted. This is the, so-called, non-verbal communication, which almost every human understands, despite language or cultural differences, without even knowing how it does.

The ever increasing human population, and the urge to fulfill basic needs, like health, well-being, security and entertainment, is leading to the creation of computer algorithms that can act like the trained professionals aforementioned, and infer from a video stream, both live and offline, the intentions on the behavior of either a single or a group of humans. This is a widely accepted prediction that computing will move further to understand human behavior and intentions in order to build *human-centered* models that could react and emulate to human affective and social signaling interactions in everyday living spaces [264]. However, several questions arises if we think on how computers are being taught to read human's non-verbal motion, actions, expressions, and intentions. How can one teach something that is empirical? Moreover, how does one teaches a computer to see? Although both the rods and cones and the camera's CCD transform light in

electrical pulses, the resemblance between our vision and computer vision stops there. The complete vision processing at brain level is something that it is not fully understood. How the electrical pulses that reach the brain rationalized in behavior is still a mystery. So can one mimics something that one does not fully understands?Researchers all over the world show that yes, it is possible, and that is the beauty of being human.

Historically, research on visual human activity recognition has been looking to mimic human perception, at least what we know, by focusing on solving the stratified problems that shaped its complex network of concepts, technical issues and dependency-connections among them. Generalizing, a human activity analysis system is comprised by three major components: *feature representation and modelling*, *detection and classification*, and *prediction and association*. These components coexist into a hierarchical and complex arrangement that represents and models spatial and temporal structures of motion, segments and detects consistent streams, and creates models for learning and inferring states and concepts.

The study of the human visual system plays a fundamental role to understand how the information may be represented and modelled by computer systems. The outstanding work of Hubel and Wiesel [136] enabled to commence the understanding of the information processing in the visual system. In the 1980s, Marr and Hildreth [211] presented one of the best known edge detectors in digital image. Later, Lowe [203] focused on the detection and description of image features, while Biederman [35] presented a theory about the arrangement of simple geometric components. Representations for activity analysis must be able to combine both capturing of distinctive properties of an object and generalization of the object's description in the presence of unknown new instances. The representation should also include cumulative and temporal information about the object [110], in order to facilitate its further detection. However, segmentation of the video stream is necessary before applying detection. But this raises the problem of a generic semantics that identify the parts of interest for an appropriate segmentation [377]. Therefore, for human action recognition a model is required to detect and segment instances of actions through the complete observation of the video stream. To close the loop around the three major components aforementioned, the inclusion of contextual knowledge in terms of how the represented objects of interest should behave and/or in terms of the inter-relational features among the objects, could improve the overall human activity recognition system in several directions, such as increasing the effectiveness of search in the scene, improvement of the learning model, and, consequently, and enhancement of the classification.

Apart from the chain of complex and inter-related steps that a visual human activity recognition outlines, we could identify, by literature overview, that the recognition of human activity can be performed at various levels of abstraction [3, 229, 377]. The most known surveys of human activity analysis in computer vision define taxonomies based by the level of abstraction from which the recognition of movement is performed, and correspond roughly to low-level, mid-level and high-level vision tasks [229]. Some of them follow instead a functional [228] or a technical-based [3] taxonomy approach. What is common from the vast majority of them, is that they describe application-based approaches for each level, revealing the importance of the application

requirements and constraints to address and evaluate human-activity-related tasks in the field of computer vision.

One of the key challenges in building automatic visual systems that can recognize human activities is the prevailing gap between the low level perceptual inputs such as pixel values or structural features, and some of the higher level inferences such as discovery of common paths, identification of individuals within a social group, classification of individuals and collective human activities, event detection, or classification of body expressiveness with context-knowledge. This semantic void is one of the main reasons for information uncertainty, which results in poor inference accuracy. We have notice that, despite the efforts in the literature, either the proposed taxonomies are semantic-based and cross-application, or functional-based and standalone application. Such division implies drawbacks not only in understanding the specific content of each application, consequently, in the technical solution, but also in the scalability of solutions for specific applications that may share concepts with applications from a different domain settings. Neither of them provides an intermediate characterization that provides a natural bridge between the type of content of the application, i.e. the perceptual inputs, and the application needs, i.e. the inferences. In this way, we propose a categorization of the analysis of human activity defined by the domain settings in which the application resides. Our work spans through three domain settings, namely *surveillance*, *multimedia* and *behavioral*, and each one provides the appropriate constraints and characteristics to investigate three levels of action categorization here defined by *the group*, *the whole* and *the parts*. Our work focuses on motion representations supported by relational-context features for the analysis of human activity at the different levels identified.

1.1 Background and Motivation

At the beginning of this thesis the aim was to provide contributions in human activity analysis at the application level. Several applications were identified along with their requirements. While this process was taking place, an extensive and intensive literature review was conducted. Considering the wide range of domains that human activity analysis covers, and the large number of works that have been reporting through almost 40 decades, this was an enormous effort, but worth it. Under our assumptions, explained next in this Section, we have identified three important domain settings, *surveillance*, *multimedia* and *behavioral*, that support our research providing the context and the research gaps to tackle.

For the last years, the vision community has been focused on developing fully automated surveillance and monitoring systems [60, 134, 173]. Such systems have the advantage of providing continuous 24 hour active warning capabilities and are especially useful in the areas of security and safety such as law enforcement, national defence, border control and airport security. The current systems perform satisfactorily while handling common issues such as illumination changes, shadows, short-term occlusions, weather conditions, and noise [60].

Generally, visual surveillance systems handle low-level vision problems, such as detection and tracking of pedestrians, and generate reports based on statistical analysis of motion representations, such as the most common paths of pedestrians [156], and warnings/notifications regarding abnormal events [191]. However, the difficulty remains in finding a generic and robust approach that can cope with the high variability of surveillance scenarios, not only related to the recording and viewpoint conditions, but also, and mainly, with the content of the scene, namely the number of pedestrians, randomness of movements, scene cluttering and scene layout. The lack of a general motion representation that can help to overcome this problem is still an open question.

Surveillance systems have been also modeling activities ranging from simple, periodic activities such as walking and running [6] to more complex ones that involve an underlying semantic structure [10,141]. During the recent years, applied research in this domain has been rapidly shifting its interest to the recognition of collective activity in public spaces. Most of them conduct their studies in crowd scenarios to solve technical problems such as detection of individuals [48], tracking [33], and detection of abnormal crowd behavior [223], among others.

Modeling social behaviors of people is an important topic to represent group activity. For this reason, surveillance applications are increasingly gaining interest on modeling individual and collective activities within a sociological point of view, since it brings more benefits to clients in areas such as retail, sports, security and safety, and smart cities. Some works have been presented recently that deal with social and physical rules to model individual's behavior within a group [126, 189, 270]. However, none of those studies aims to understand the group's behavior as a social entity, modelling the relationships between the individuals within the same group, and the relationships between the individuals and the scene. Other works investigate the addition of contextual information to improve the group activity recognition performance [65, 181, 182]. However, they model the motion features and relationships depending on the activities of interest. No interpretation about the sociological meaning of relational features, such as position-based and attention-based, and social context is outlined by any work in the computer vision literature, to the best of our knowledge.

Modeling human action and activity patterns for recognition or detection of special events has been attracting significant research interest in recent years. Spatio-temporal motion representations are mandatory to tackle this challenge. Multimedia event retrieval is one of the areas that has been facing an exponential interest on this type of processing [201,335]. The task involves the identification of complex actions and events in unconstrained videos with different conditions of camera motion and viewpoint, human editing, and background clutter. In this scenario, an event is not just a sudden change of conditions, as normaly is conceptualized in surveillance settings, it is a semantic abstraction of a large activity composed by one or more concepts such as objects, actions and scenes.

The most robust approaches for the detection of complex event from multimedia videos are the ones that include complementary multimodal features, such as motion representations [246,371], appearance scene representations [258], object detection [351], and others. Such features are denominated by *hand-craft visual features* and are being integrated with more discriminative Con-

volutional Neural Network (CNN) descriptors [142, 395]. Different encoding and aggregation methods of features have been tested in the literature [103, 145, 259]. Despite the increasing attention for CNNs approaches, the most successful methods for event detection are still aggregating *hand-craft visual features* [103].

Human activity analysis also plays an important role in gathering meaningful information to understand human behavior in social relationships. Computer vision research on Human Behavior Analysis (HBA) includes a broad range of studies on developing computer systems and models to achieve nonverbal sensitivity in different contexts and through different channels such as face, voice, gait and body gesture [81, 298, 386]. Context of a visual environment is essential to properly interpret behavior in terms of the functionality of the object and in the intention involved in the action. The pertinence of developing automatic computer systems and models to achieve nonverbal sensitivity in different contexts is increasing [264]. Several studies in psychology have indicated important remarks: the combination of facial expression and body gesture are more informative than each alone [221]; gesture expressiveness is transmitted unconsciously [124]; body expressive cues such as head inclination, shifting posture, and face touching are often attached to social affective states [91]; postures affirm the current attitude of people towards social situations [322].

Nonverbal Communication Computing (NCC) has been widely investigating the aforementioned topics employing motion analysis methods. However, no study has analyzed the variation of body expressiveness, as a whole, in a duo-interaction conversational scene, and distinguished motion expressiveness patterns among different emotional contexts. Such type of study could enable the creation of nonverbal behavioral models in social interaction environments useful to develop assistive technologies to support the deficits of specific populations.

Outlying the literature review, we highlighted three important conclusions that represent the foundation of our work. First, the importance of motion behind any kind of human activity is crucial, and its combination with contextual information either in terms of scene knowledge, or relational features among objects of interest, is even more important due to their complementarity. In this way, several works and semantics that use motion and context in some way were identified. Second, relating both sources of information, we found a gap between the definition of solutions for specific applications and the semantics associated to common human-activity-related tasks. On one hand, the structure of the proposed semantics in the literature normally follows a hierarchy of dependent and complex entities related to human motion, such as *action primitive*, *action* and *activity*. Them different methods or techniques are used to represent and model each semantic level. However, such semantics may not be useful or possible to identify in other application domains. On the other hand, some application do not permit to implement solutions that may share semantics, or representations, with other applications, even in the same domain, due, for instance, to restricted requirements. We believe that taking into consideration the domain settings in which the application resides, a level of activity can be defined within which motion and contextual representations can be investigated for the analysis of human activity. This categorization has two advantages: i) more applications in the same domain settings may benefit from the representations and semantics proposed; ii) similar semantics structures, already defined in the literature, that

resides in different domain settings may benefit of the analysis from different levels. Last but not least, we stated that most of the research in this area is application-oriented at some point, or by definition or by evaluation. From the aforementioned conclusions, we may outline the overall context of our thesis in Fig. . From top to bottom, it defines relationships among the more abstract representations, passing through the definition of the different levels of human activity and their correspondence with a domain settings, that defines the research questions to address. Finally, the applications define the functional component of our contributions, presented in terms of frameworks.

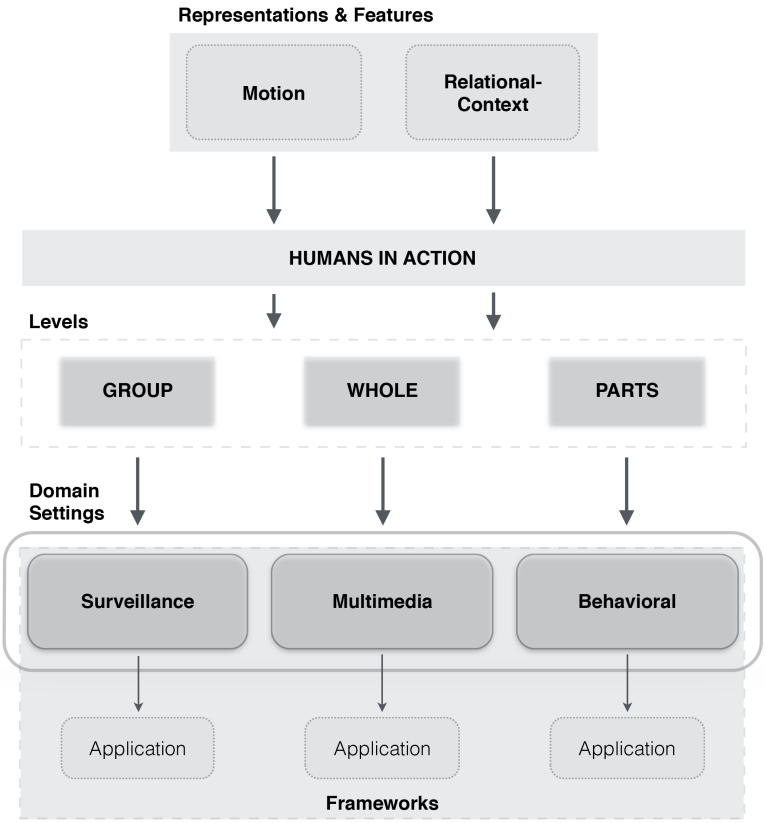


Figure 1.1: Overall context of the Thesis.

1.2 Goals

We aim at investigating motion and relational-context representations to support the modelling and detection of human actions at three different levels: i) *the group* in the scene, where global motion patterns and social collective activity are analyzed; ii) *the whole* in the frame, where motion provides evidence of relevant foreground and different kinematics and contextual features are integrated for classification; iii) *the parts* in the body, where segmented motion regions are represented to characterize expressiveness.

This thesis conducts its research methodology based on two premises: the need to address *research questions in real-world scenarios*, and the development of *complete frameworks*. We argue that research should be performed and evaluated under real conditions and should tackle real problems. We also emphasize the importance of a complete framework, specially in the field of human activity analysis, where many processing modules are involved until the final inference stage. Both factors permit to reduce the gap between the academic and industry fields, while providing valuable and unique conditions to improve research methodologies and outcomes. The identified domain settings were selected taking into consideration their unique characteristics to address each level of human activity. However, they were also selected taking into consideration their increasing attention by the academia and industry.

This thesis proposes four frameworks within three different domain settings. All of them are evaluated under an application-based perspective. The aim of each one is to provide a solution for the specific application needs and, mainly, to propose new approaches or improvements in the literature, in previously identified gaps or in incomplete solutions, respectively. In this way, we have identified the following global challenges: i) efficient global motion representation for different surveillance scenarios; ii) individual and collective representations for social behavior analysis in surveillance scenarios; iii) importance of motion to identify relevant movement and integrate contextual information into multimedia video classification; iv) combination of motion features that capture expressiveness intention and characterization in non-verbal communication scenario.

The proposed representations on each domain settings deal with the specific application-constraints, and do not necessarily mean that they are applicable to the remaining settings handled in this thesis. Table 1.1 summarizes the specific goals of this dissertation, indicating the respective domain settings and the proposed framework to which it belongs.

Settings	Surveillance - Motion patterns detection.
Framework	Motion analysis.
Research Goal	Provide long-range motion trajectories.
Settings	Surveillance - Social behavior understanding.
Framework	Relational features analysis.
Research Goal	Classification of individual and collective behavior, and provide evidence of the sociological meaning of each feature.
Settings	Multimedia - Video classification.
Framework	Multi-modal visual analysis.
Research Goal	Provide relevant spatio-temporal representations based on motion evidence.
Settings	Behavioral - Expressiveness characterization.
Framework	Body expressiveness analysis.
Research Goal	Provide a discriminative combination of motion features to distinguish motion expressiveness patterns among different emotional contexts.

Table 1.1: Summarized goals of this thesis organized by domain settings.

1.3 Thesis' Scope

This work is largely based on motion as main driving force of human action. This foundation is based on an extensive process of literature review, which states that since the early 1970s, when human activity analysis became a field of study, motion assumed a crucial role inspired by the experiment of the moving light display (MLD) of Johansson [154]. The first surveys in the computer vision field about human activity analysis focused mostly on taxonomies and techniques related to motion extraction and articulated nonrigid motion [4–6, 59, 105]. The most recent survey of Afsar *et al.* [2] presents a review from 2000 to 2014, where 193 papers are classified by detection techniques, datasets and applications. Under the techniques' category, the papers are analyzed following the taxonomy defined by Moeslund *et al.* [229] (initialization, tracking, pose, recognition). This means that motion is considered within the study of all the reviewed papers related to human activity analysis, which clearly states its importance.

Motion detection permits to distinguish moving objects from background or moving relevant entities from irrelevant entities, depending on the abstraction level. Subsequent processes such as tracking and activity recognition are greatly dependent on it. The process of motion detection may involve several subprocesses depending on the application, such as motion camera compensation, environment modeling, motion segmentation, and object classification. Any of these subprocesses may intersect each other during the whole processing. This complexity expressed by the various processing steps and their interconnectivity and dependence, suggests the creation of frameworks or systems capable to provide a valid solution that incorporates reliable features, discriminative representations and robust inference methods, in order to reduce the level of uncertainty associated with the interpretation of human behavior.

Since its birth, automatic recognition of human activities has been a hot trend in the research community of computer vision and artificial intelligence, as well as in the industry. However, nowadays there are two factors in the society that have been promoting exponentially the interest and demand for this field, namely the *large deployment of small electronic devices with an embedded camera*, which democratizes the production of multimedia content and *terrorism*, which has lead, again, to extreme security measures worldwide. The former implies a huge production of content that is constantly uploaded to the internet using multimedia platforms like Facebook [1] and YouTube [2], which requires automatic solutions to inspect and retrieve information from pre-defined labels in terms of complex events. The latter involves two factors: firstly, the claim for algorithms that automatically recognize complex individual and collective human activities and identify suspicious events in surveillance streams; secondly, the need to enlarge the forensic technologies with robust solutions that identify nonverbal cues in a context-sensitive manner.

This dissertation covers the analysis of motion representations and relationships among motion entities, and take into consideration the influence of context under three domain settings, namely *surveillance*, *multimedia* and *behavioral*. The considered context can be scene-based, i.e.

[1] http://www.facebook.com
[2] http://www.youtube.com

physical, or sociological-based, i.e. venue or conversational topics. The methodology of the conducted research is application-based, which leads to contributions in four specific topics, *motion representations and patterns detection*, *social behavior analysis*, *video classification* and *body expressiveness analysis*. Fig. 1.2 outlines the line of work of this thesis, highlighting the proposed frameworks for each covered topic, the specific contributions in each one, and the respective application domain.

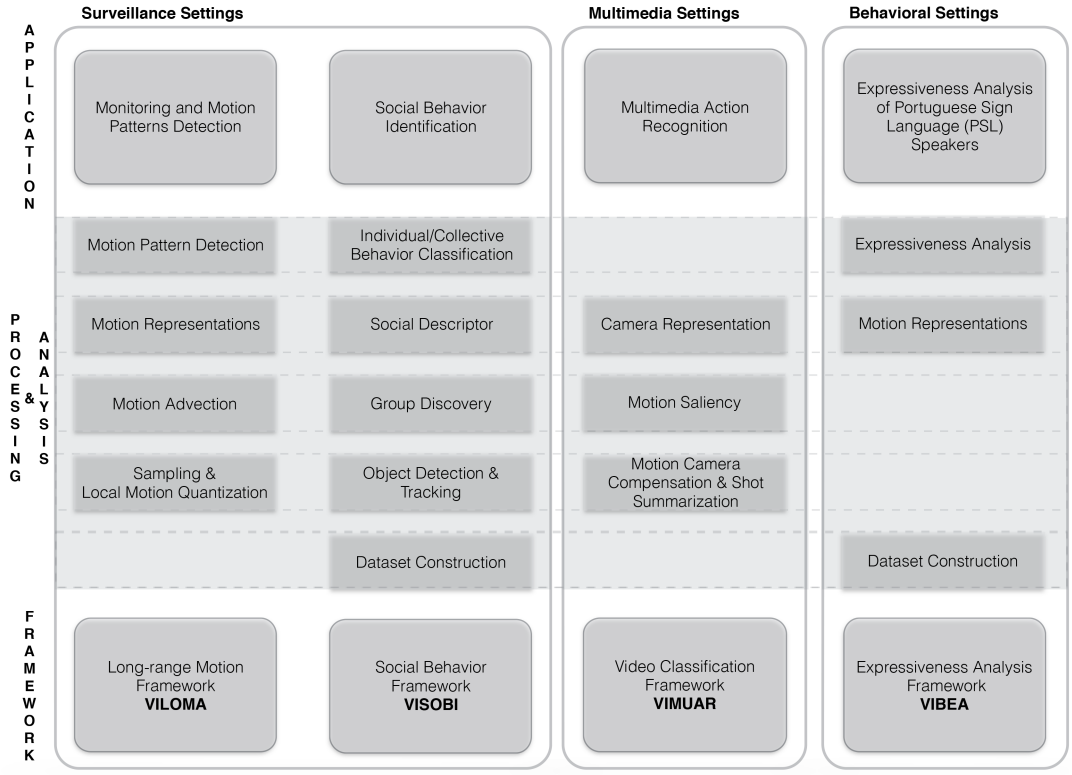


Figure 1.2: Line of work of the Thesis.

1.4 Thesis' Structure

This thesis is organized into seven Chapters, each one describing the work led during the last four years. Given the broad scope of topics covered in this thesis, each Chapter follows a common body structure to facilitate reading: i) *Introduction*, refers to a summary of the background and motivation for the current study; ii) *Overview*, outlines the state-of-the-art of some specific and relevant topics for the current study; iii) *Validation*, reports the evaluation methodology and the experiments details, highlighting the main critical results; iv) *Summary*, conveys a final overview of the current problematic, outlines the presented solution with the main contributions, and stresses the main conclusions. Each Chapter is also accompanied of one or more Sections that describe the characteristics of the problem and of the defined approach along with its novel contributions. This Chapter presents the background and main motivations, frames the thesis' line and the adopted

methodology, and outlines the goals and original contributions of the thesis. The remainder of the dissertation comprises the following chapters.

In Chapter 2 a literature review is provided. It starts with an overview of the origins and progress of the vast field of human activity analysis, highlighting the main taxonomies used along the last two decades. Afterwards, the three application settings that this thesis covers are discussed, namely vision surveillance, multimedia video, and human behavioral analysis. In each one, more specific topics, which work as a basis for the developed work, are surveyed: i) in Surveillance, motion representations, patterns detection and identification approaches, and social behavior analysis are enumerated; ii) in Mutimedia, spatio-temporal mid-level representations and techniques for complex event detection are exposed; iii) in Behavioral, psychology-related works, multi-features visual representations and frameworks to understand body behavior are reviewed. Finally, a summary, open-questions and near future trends about human activity analysis in general are discussed.

Chapter 3 presents a motion flow framework with the main goal to robustly obtain long-range trajectories and several spatio-temporal motion representations in different surveillance scenarios. For such generic framework and considering the multidisciplinary theoretical concepts involved, a specific survey is taken to encompass microscopic and macroscopic approaches, Lagrangian and Eulerian flow descriptions, streaklines and streamlines representations, vector field representation and motion advection. Due to the novelty of the proposed work, a new evaluation methodology is designed, presented and discussed. Advantages of the framework in human activity-related tasks are evaluated, namely motion patterns segmentation.

In Chapter 4 a novel approach to bring a sociological perspective into the analysis of human activity in visual surveillance is elaborated. A novel dataset with low-level and high-level annotations is presented. The proposed activity-related semantics are explained. The complete framework for analysis of individual and collective behavior is described considering the automatic extraction of features, including a group discovery algorithm, the formulation of the social descriptor, the techniques for feature relevance analysis, and a dual-classification pipeline. The impact of the automatic features extraction in the final classification results are reported, and a sociological analysis of the individual features meaning is approach.

Chapter 5 describes the proposed video classification framework that uses motion to detect relevant foreground components, acquire camera context and represent foreground with complementary descriptors. The most important modules of the framework are surveyed, the individual contributions highlighted and their intuition explained. Several quantitative evaluations are presented for each processing modules, regarding their performance in supplementary tasks and/or their impact in the final classification. Finally, a comparison with the state-of-the-art is reported.

A study of body expressiveness analysis in social context is presented in Chapter 6. A related survey is carried on low-level representations such as trajectory-based and pixel-based, and on mid-level representations like body gesture, expressive gesture, and body expressiveness. A novel dataset of duo-interaction between Portuguese Sign Language (PSL) speakers is presented, along with all the related requirements, constraints, protocol, physical setup and final characteristics.

Three main research questions are traced and the corresponding methodologies are described.

The dissertation is summarized in Chapter 7. Final conclusions and possible future lines of work, including improvements to the solutions proposed in the previous chapters, are described.

Additionally, an Appendix A provides auxiliary material to complement the study presented in Chapter 6.

1.5 Contributions

Considering the line of work followed in the course of this thesis, and illustrated in Fig. 1.2, we summarize below the contributions of our work for the representation and analysis of human activity at different levels.

1. **Long-range motion analysis**, considers the proposal of a generic and dynamic motion analysis framework capable to automatically extract valuable representations for human activity-related tasks on several types of surveillance scenarios (see Chapter 3), with the following characteristics:

   i) new global outlier removal technique for flow vector data;

   ii) fine-to-coarse global representation of flow vectors;

   iii) integration of local motion information, based on information theory principles, with global motion information, based on temporal integration of flow, to capture longer spatial and temporal changes in the scene;

   iv) re-correlation algorithm to link broken streamlines and accurately form long-range streamlines which correspond to the global long-range motion trajectories.

2. **Individual and collective social behavior analysis**, considers the proposal of a complete and automatic framework that identify Individual Profiles (I.P.) and Group Behaviors (G.B.) considering relationships among the individuals and with the scene context (see Chapter 4), with the following characteristics:

   i) novel dataset with low-level and high-level sociological annotations;

   ii) new semantics for I.P.s and G.B.s;

   iii) automatic features extraction, including a proposed group discovery algorithm;

   iv) multi-resolution descriptor that identify and encode social interactions cues as features;

   v) dynamic classification approach based on mini-batches.

   vi) kernel-based feature relevance analysis approach using a center kernel alignment criteria.

3. **Motion evidence for spatio-temporal representations**, considers the proposal of a video classification framework that extends and improves individual components to increase the classification results (see Chapter 5), with the following characteristics:

   i) robust camera motion compensation and characterization technique;

   ii) new camera motion descriptor to express action-related context;

   iii) combination of motion with foveal-camera information to build a foreground saliency map;

   iv) shot-based approach for video classification.

4. **Expressiveness body analysis**, considers the proposal of a framework that extract and combine meaningful motion-related features to identify various motion expressiveness patterns under different emotional contexts (see Chapter 6), with the following characteristics:

   i) novel dataset of a duo-interaction between deaf and hearing people with different conversational topics;

   ii) extraction and integration of discriminative and complementary motion features;

   iii) classification methodology to differentiate deaf from hearing people;

   iv) classification methodology to differentiate conversational topics;

   v) discovery of levels of mastery in PSL from visual and user-survey information;

List of Publications Related to the Thesis

The work related with this thesis resulted in the publication of the followings journal papers:

- E. M. Pereira, S. Molina-Giraldo, L. Ciobanu, D. Insuasti-Ceballos, A. Álvarez-Meza, G. Castellanos-Dominguez, and J. S. Cardoso. Group Discovery and Feature Relevance Analysis for Social Behavior Identification. IEEE Transactions on Circuits and Systems for Video Technology, 2015 (Submitted)
- E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Cross-layer Classification Framework for Automatic Social Behaviour Analysis in Surveillance Scenario. Journal of Neurocomputing, 2015
- E. M. Pereira, J. S. Cardoso, and R. Morla. Long-Range Trajectories from Global and Local Motion Representations. Journal of Visual Communication and Image Representation, 2015

From this thesis resulted the following international conference papers:

- M. Khodabandeh, S. Muralidharan, A. Vahdat, N. Mehrasa, E. M. Pereira, S. Satoh, G. Mori. Unsupervised Learning of Supervoxel Embeddings for Video Segmentation. ICPR - Iternational Conference on Pattern Recognition, 2016
- E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Social Signaling Descriptor for Group Behavior Analysis. In Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, pages 13-22, 2015
- I. Rodrigues, E. M. Pereira, and L. F. Teixeira. Analysis of Expressiveness of Portuguese Sign Language Speakers. In Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, pages 708-717, 2015
- E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Context-based Trajectory Descriptor for Human Activity Profiling. In Proceedings of The IEEE International Conference on Systems, Man and Cybernetics, pages 2385-2390, 2014
- E. M. Pereira, J. S. Cardoso, and R. Morla. Motion Flow Tracking in Unconstrained Videos for Retail Scenario. In Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, pages 340-349, 2013

The following technical reports also derived from this thesis:

- E. M. Pereira, J. S. Cardoso, and R. Morla. Long-Range Trajectories from Global and Local Motion Representations. In arXiv (http://arxiv.org/abs/1509.08647), 2015
- S. Phan, D. D. Le, S. Satoh, E. M. Pereira, S. Barnwal, M. Klinkigt, Y. Watanabe, and A. Hiroike. Multimedia Event Detection system in TRECVID 2015, 2015

I also participated in national conferences with the following papers:

- E.M. Pereira, and J. S. Cardoso. Analysis of Pedestrian Models for Social Behavior Discovery in Surveillance Scenario. In Proceedings of the 21^{th} Portuguese Conference on Pattern Recognition, 2015
- E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Social Descriptor for Individual Profiling and Group Behaviour Analysis. In Proceedings of the 1^{st} Doctoral Congress in Engineering, 2015
- E.M. Pereira, J. S. Cardoso, and R. Morla. A Critical Analysis about a Motion-based Approach to Extract Global Trajectories. In Proceedings of the 19^{th} Portuguese Conference on Pattern Recognition, 2013

- E.M. Pereira, J. S. Cardoso, and R. Morla. Motion Flow Trajectories: The Beginning of a Visual Perceptual Reasoning. In Proceedings of the 18^{th} Portuguese Conference on Pattern Recognition, 2012

The following publications result from work partially related with this thesis that was not included in this document:

- A. S. Domingues, F. Barbosa, E. M. Pereira, M. B. Santos, A. Seixas, J. Vilas-Boas, J. Gabriel, and R. Vardasca. Book Chapter, Infrared Thermography in Swimming: Thermal Characterization of Swimming Techniques. In IGI-Global Innovative Research in Thermal Imaging for Biology and Medicine, 2016
- A. S. Domingues, F. Barbosa, E. M. Pereira, M. B. Santos, A. Seixas, J. Vilas-Boas, J. Gabriel, and R. Vardasca. Towards a Detailed Anthropometric Body Characterization using the Microsoft Kinect. In Proceedings of Technology and Health Care Journal, 2015
- A. S. Domingues, E. M. Pereira, J. Gabriel, and R. Vardasca. Case Study in Thermal Monitoring of Physiotherapy Treatments to Ankle Sprains in Rugby Athletes. In Proceedings of Pan American Journal of Medical Thermology, 2014

Chapter 2

Literature Review

> One can have no smaller or greater mastery than mastery of oneself.
>
> Leonardo Da Vinci

Vision-based human activity recognition has been facing an increasing importance among the computer vision and artificial intelligence communities with applications in visual surveillance, multimedia video retrieval and human behavior analysis. In general terms, the developed systems so far have been focused on unveiling features, representations and interpretations around four levels of complexity *movement*, *action*, *activity*, and *behavior*. This hierarchy of concepts encompasses the analysis of the relationships between abstract entities, influence of context, and meaning of descriptive semantics. The research efforts aim to include and model motion components, objects, people, interactions and scene to reduce the uncertainty and errors associated to the overall interpretation of human activity, while providing evidence of the contributions of each entity to properly understand semantic and social behavior.

2.1 Human Activity Analysis at a Glance

Human activity definitely implies motion as its main characteristic. One of the first researches into the nature of human motion was conducted by the photographers E. J. Marey and E. Muybridge in the 1850s. They photographed moving subjects and revealed several interesting and artistic aspects involved in human and animal locomotion. The classic moving light display (MLD) experiment of Johansson [154] provided a great impulse to the study and analysis of human motion perception in the field of neuroscience, and it was an enormous inspiration for the early works about human activity analysis in the computer vision field.

Human activity recognition is an important area of computer vision research that has been facing an increasing interest in the research and industry fields during the last years, thanks to the advances of information technology [2], among other factors. It is a multidisciplinary area that involves diverse fields such as psychology, biomechanics, artificial intelligence, and pattern recognition. Such a variety of knowledge implies a subdivision into several research subtopics

such as analysis of human motion capture [227], behavioral biometrics [398], visual video surveillance [173], and human behavior detection [110], just to mention a few.

From a technical point of view, vision-based human activity recognition is the process to automatically analyze and recognize the ongoing activities from an unknown video [3]. The video could have one or more activities, and the challenge could be not only to correctly classify the video into its activity category, but also to detect the starting and ending frames of the continuous activities within the video. The video-content may represent diverse scenarios such as surveillance streams, multimedia videos, and behavioral recordings, and their visual processing may be useful for various applications such as detection of common motion patterns in visual surveillance, multimedia video classification and body expressiveness analysis, respectively.

Human activity recognition is challenging due to variations in recording settings, motion performance, non-rigid movement, self and mutual-occlusions, inter-personal differences, and segmentation of changing scenes in natural and uncontrolled environments [294]. To overcome these challenges, visual analysis of human activity focuses on three general functions [110]:

i) *Representation and Modelling*, to extract and encode visual information from imagery data in a more concise form in order to capture the intrinsic characteristics of the objects of interest. This step encompasses low-level feature extraction and mid-level representation processes;

ii) *Detection and Classification*, to discover and search for salient and unique characteristics of certain object behavior patterns from a large quantity of visual observations, and to discriminate them against known categories. This step encloses learning and interpretation of high-level semantic concepts;

iii) *Prediction and Association*, to forecast future events based on the past and current interpretation of behavior patterns. This step intents to achieve concept identification through behavioral expectation and trend.

One of the main issues within the human activity analysis field is the taxonomy approach, mostly defined by disciplines such as computer graphics, biomechanics, and psychology [173, 294]. Normally, the diversity of taxonomies is based by the level of abstraction from which the recognition of movement is performed. Concepts like *gestures*, *actions*, *simple actions*, *complex actions*, *activities*, *group activities*, *nonverbal behavior* are often used interchangeably in the literature. For instance, Nagel [238] suggests a hierarchy of action based on *change*, *event*, *episode* and *history*. Bobick [39] proposes different levels of abstraction for motion, *movement*, *activity* and *action*. González *et al.* [111] includes the term *situations*, and Jenkins *et al.* [146] applies a hierarchy of *action primitives* and *parent behaviors*. The complete definition of a taxonomy permits to describe the concepts adequately and provides a base to create a common terminology for comparison purposes.

Some surveys target their analysis based on the selection of a taxonomy. For instance, Moeslund *et al.* [229] proposed a hierarchy based on *action primitive*, *action* and *activity*, where an *action primitive* is an atomic movement that can be described at the limb level; an *action* consists of action primitives and describes a whole-body movement, that might be cyclic; and an *activity* contains a number of consecutive actions, which gives an interpretation of the movement that is

being performed. Aggarwal and Ryoo [3] follow an approach-based taxonomy divided into *single-layered* and *hierarchical*. Such categorization is based on the concept of *complexity* of the human activity, which defines four levels: *gestures*, *actions*, *interactions*, and *group activities*, where *gestures* are the atomic components that describe the movements of a human's body part; *actions* are single person activities composed of various sequential *gestures*; *interactions* are human activities that involve two or more persons and/or objects; and *group activities* are activities performed by groups with multiple persons and/or objects. Turaga *et al.* [354] also base their analysis in *complexity*, but they only consider high-level recognition at two levels, *actions* and *activities*. These taxonomies present different granularity levels to represent the problem of human activity analysis, however, all of them consider at least one level for context such as the environment and/or the interaction between persons and/or objects.

Other taxonomies presented in the literature guide their approach by technical considerations. In particular, Moeslund and Granum [228] follow a functional taxonomy categorized by four groups: i) *initialization*: to ensure that the system starts with a correct interpretation of the scene; ii) *tracking*: to segment the objects of interest from the background and match correspondences between consecutive frames; iii) *pose*: to estimate the pose in corresponding frames; iv) *recognition*: to recognize the behavior, identify the individuals, and the actions of an individual or a group. More recently, Afsar *et al.* [2] follow the same taxonomic designation, but also include high-level methods for human behavior recognition, while at the same time compare distinct methods in each group stage, pointing out advantages and limitations. The work of Poppe [294] only considers full-body movements, it does not explicitly considers context and interaction between persons and/or objects. His analysis is merely focused in low-level image representations for action recognition. Aggarwal and Cai [4] present a complete action recognition system based on three global steps: extraction of human body structure from images, tracking across frames, and action recognition. Wang *et al.* [370] use a similar taxonomy based on human detection, tracking, and behavior understanding. Wang and Singh [369] divide human movement into tracking and motion analysis, where tracking is discussed for full body and some body parts such as hands and head. Cedras and Shah [59] argue that motion is a more important cue for action recognition than the structure of the human body, therefore they introduce a survey on motion-based approaches to recognition as opposed to structure-based approaches. Gavrila [105] focuses his analysis mainly on the tracking of hands and humans via 2-D or 3-D models and presents a discussion of action recognition techniques. However, these types of analysis have been shifting more toward recognizing actions from tracked motion or structure features and on recognizing complex activities and their behavioral semantics in real-world settings. Fig. 2.1 outlines the general three levels of representations that compose a human activity analysis system, accompanied with the most common processing modules and concepts of taxonomies surveyed in the literature.

Due to an increasing advance of the information technology, computational frameworks have been emerging for the automatic detection of human behavior in real-world settings, which permit to explore higher levels of abstraction previously studied only in controlled environments by disciplines such as social signaling and human computer interaction (HCI). This has led to the

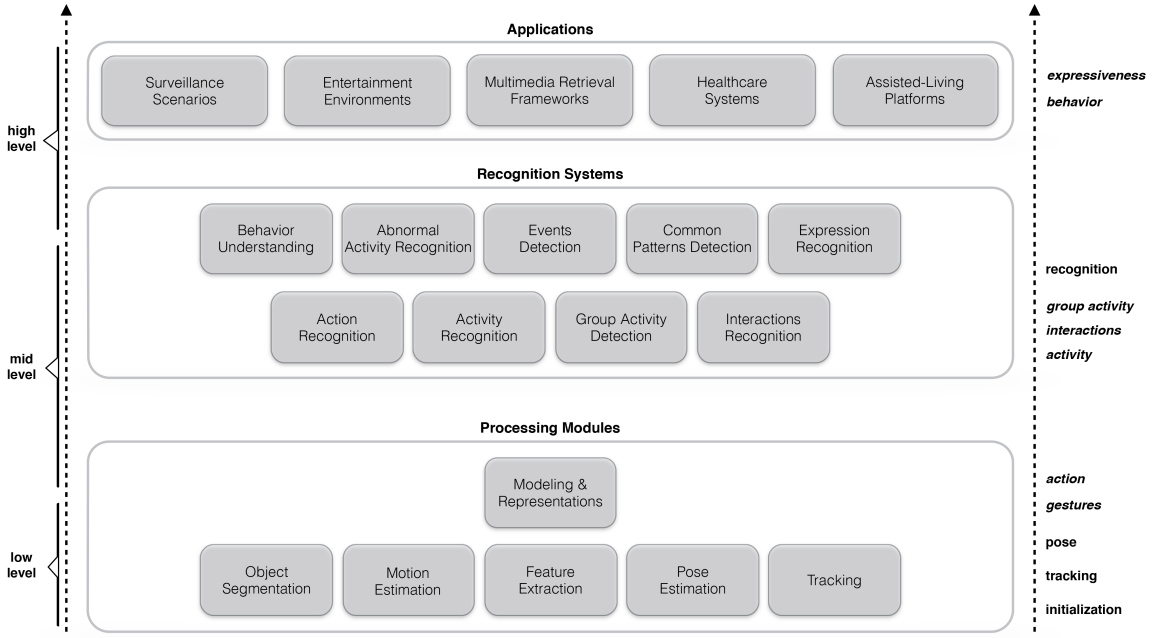


Figure 2.1: Overview of a general system for human activity analysis. The most common concepts from the taxonomies exposed in the literature are also presented in a bottom-up perspective accordingly with their definitions and with the necessary technical components to accomplish their detection.

enlargement of concepts covered by the field of human activity analysis, allowing the interest for social and expressiveness characteristics of nonverbal behavior. The intersection of concepts and interests favour the interchangeability of knowledge among disciplines. On the one hand, deeper insights into nonverbal communication computing provide new models for perception, cognition and learning of action, interaction, and groups at social and behavioral levels [225]. On the other hand, a large variety of methods to detect and track motion permits to include useful features into multimodal nonverbal frameworks for expression recognition, behavior analysis, sign language recognition, group activity recognition, among others. Group activity detection is one of the fields that has been shortening the frontiers among human activity analysis and nonverbal communication computing. Surveillance applications are increasingly gaining interest on modeling individual and collective activities within a sociological point of view, since it brings more benefits to end-users in areas such as retail, sports, security and safety, and smart cities.

Table 2.1 shows some of the most relevant surveys in the area of human activity analysis for the last three decades. Considering the aforementioned interests of social behavior, we also include some works that deal with nonverbal behavior understanding. The borderline between those works and the ones that frame into the Social Signal Processing (SSP) is very narrow, and such separation is still not very clear in the literature [360]. In this thesis, we make the distinction considering two factors: *application-based*, where social-aware applications [275] have a great interest by the SSP field; and *social signs and behaviors to identify*, where social cues that convey information about

feelings, mental state, personality, and other traits of people [305], fall inside the SSP field. Such works are out of scope for this thesis.

Year	Authors	Focus
1994	Aggarwal *et al.* [5]	Articulated and elastic nonrigid motion.
1994	Cedras and Shah [59]	Motion extraction.
1995	Aggarwal *et al.* [6]	Articulated and elastic nonrigid motion.
1997	Aggarwal and Cai [4]	Motion extraction.
1997	Gavrila [105]	Motion estimation and recognition.
1997	Shah and Jain [330]	Motion recognition.
1999	Pentland [274]	Person identification, surveillance, 3D methods, and perceptual user interfaces.
2000	Moeslund and Granum [228]	Initialization, tracking, pose estimation, and recognition.
2001	Buxton [54]	Recognition.
2001	Wang *et al.* [370]	Detection, tracking, and recognition.
2001	Liang *et al.* [194]	Motion analysis.
2002	Wang and Singh [369]	Tracking and motion analysis.
2003	Hu *et al.* [134]	Surveillance.
2004	Aggarwal and Park [7]	Recognition.
2005	Valera and Velastin [355]	Distributed surveillance.
2006	Poppe [293]	Recovery human pose (modeling and estimation) and motion.
2006	Moeslund *et al.* [229]	Initialization, tracking, pose estimation, and recognition.
2006	Forsyth *et al.* [100]	Tracking and motion synthesis.
2007	Yampolskiy and Govindaraju [398]	Behavioral biometrics.
2007	Krüger *et al.* [176]	Representation, recognition, synthesis and understanding of action.
2007	Morris *et al.* [235]	Trajectory-based activity analysis for video surveillance.
2007	Zeng *et al.* [410]	Affective behavior recognition in real-world settings.
2008	Turaga *et al.* [354]	Recognition of actions and high-level activities.
2008	Weinland *et al.* [377]	Full-body motions representation, segmentation and recognition.
2008	Chandola *et al.* [62]	Anomaly detection.
2009	Poppe [294]	Recognition.
2010	Aggarwal and Ryoo [3]	Recognition of gestures, actions, interactions, and group activities.
2010	Ko [173]	Surveillance
2010	Chaaraoui *et al.* [60]	Surveillance and Ambient-assisted living.
2011	Bousmali *et al.* [41]	Nonverbal behavior analysis during displays of agreement and disagreement.
2011	Kleinsmith and Nadia [168]	Affective body expression perception and recognition.
2012	Metaxas *et al.* [225]	Motion analysis methods for recognition of nonverbal behavior.
2012	Ke *et al.* [162]	Human object segmentation, feature extraction and representation, activity detection and classification.
2012	Chaquet *et al.* [64]	Datasets for action and activity recognition.
2014	Afsar *et al.* [2]	Initialization, tracking, pose estimation, and high-level human behavior recognition.

Table 2.1: Some surveys about human activity analysis. The *Year* column corresponds to the year of the most recent paper in the survey.

In the following, we present a brief about human activity analysis in three settings: *surveillance* (Section 2.2), *multimedia* (Section 2.3) and *behavioral* (Section2.4). On each one, some important works are highlighted and subdivided into conceptual and technical aspects that are of major interest for the methodology pursued along this thesis.

2.2 Surveillance Settings

Automated visual surveillance is one of the most significant drivers for visual analysis of behavior in social environments. Normally, the computer vision solutions are built to automatically interpret human activities and detect unusual events that could be dangerous for public safety, while also assisting human operators to focus attention on more relevant threats. A general automated visual surveillance system consists in modeling the scene to detect the moving objects, classifying them, track the objects of interest, and then, through the information collected along the frames, analyze their behaviors (see Fig. 2.2). Nowadays, this type of system may work in offline mode, for instance to extract statistics for sport activities or highways/railways traffic, or in real-time, in order to provide useful information to predict and avoid unwanted events for military services or patrolling of country borders.

Further review focuses on different low-level and mid-level motion representations (Section 2.2.1), modeling and learning of such representations to detect and identify high-level patterns (Section 2.2.2), and an overview of the most recent trend in surveillance settings that ambition the understanding of social behavior within a socio-psychological point-of-view (Section 2.2.3).

2.2.1 Motion Representations

Different moving regions may correspond to different moving objects in real-world scenes. The classification and tracking stages in the surveillance system (see Fig. 2.2) depend on the type of representation of the object of interest. Cedras and Shah [59] divide the extraction of motion information into three different representations: i) *trajectory-based*: where motion trajectories, spatio-temporal curves and reference curves provide features such as velocity, speed, direction, joint angles, spatio-temporal curvature, and they can be used in recognition processes to detect motion-based and relative motion-based events; ii) *optical flow-based*: where normal flow statistics, correlation and average flow of a region are some of the features used in this type of representation; iii) *region-based*: where binary and graylevel image features can be encoded into mesh a feature codebook and model-view eigen images.

Trajectories are very popular because their interpretation is straightforward. They are based on the detection of points of interest and their correspondences in subsequent frames. In some scenes their computation can be really difficult. Also, the correspondence problem can be combinatorial demanding and affected by occlusion, noise, and periodic textures. Trajectory parametrization may offer discriminative information. For instance, Takahashi *et al.* [349] propose a human recognition method based on multiple trajectories created by detecting and tracking local key-points around moving objects. They create a motion-speed invariant feature descriptor that separates the

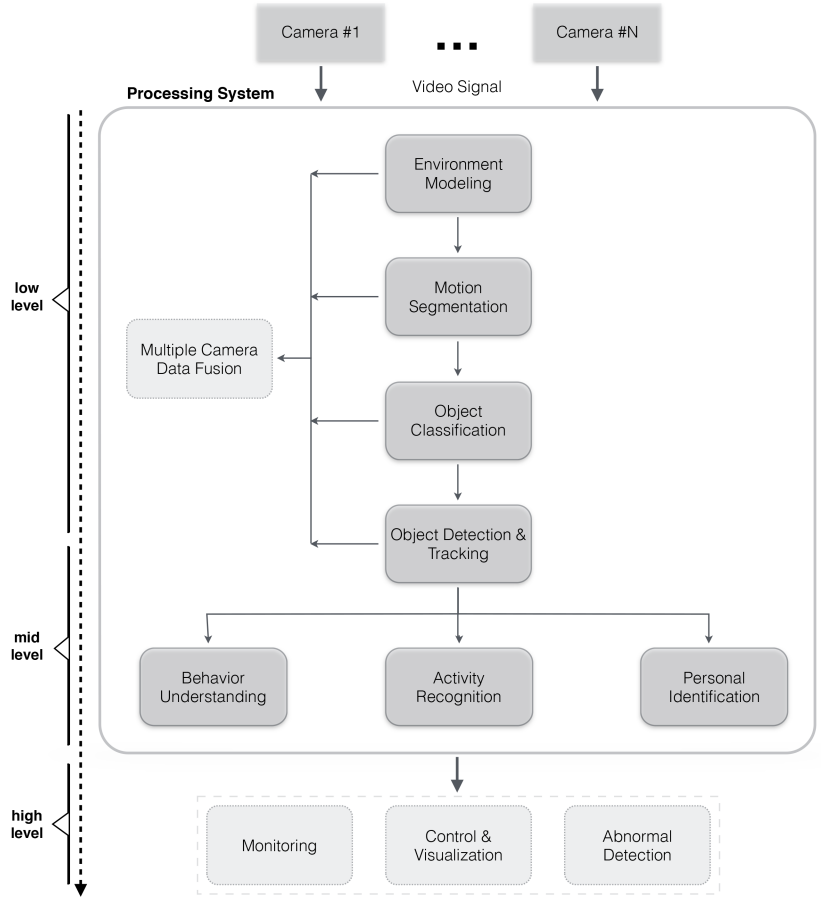


Figure 2.2: A general framework for an automated visual surveillance system.

trajectory into motion vectors, build a dual-frequency histogram from the direction and magnitude of those vectors, and finally combine it with appearance features for further clustering. Culter and Davis [73] describe a self-similarity-based technique, and its evolution over time through tracking, to analyze periodic motion from articulated objects, such as humans. Khalid and Naftel [163] consider trajectories as time-series and apply a Discrete Fourier Transform (DFT) to obtain the coefficients of the basis functions and use them as input for a Self-Organizing Map (SOM).

Optical flow-based representation has successfully been used as a source to describe coherent motion of points or features between image frames and to detect independently moving objects, even in the presence of camera motion [370]. However, it suffers from the aperture problem, it is prone to boundary oversmoothing, its computation can be computationally demanding, and is very sensitive to noise. Some of the works within this representation present low-level approaches to segment moving objects over time. For instance, Bregler [44] represents each pixel by its optical flow and describes a probabilistic decomposition of human dynamics at multiple abstractions: low-level primitives with coherent motion grouped by Expectation-Maximization (EM) clustering, mid-level categories represented by dynamical systems, high-level gestures represented by Hidden Markov Models (HMM). Rowley and Rehg [312] use an EM approach to segment the flow field

into different articulated parts, adding kinematics motion constraints to each pixel. Stringa [341] proposes an algorithm based on morphological filters for scene change detection that maintains stationary performance even in varying environmental conditions. Lipton [197] uses the residual flow to analyze rigidity and periodicity of moving objects, and classify them among vehicles or humans through a technique called *dynamic region matching*. Lin *et al.* [196] propose a novel method for modeling dynamic visual phenomena from aggregate motion fields of moving people. They extract the constituent elements of flow by a geometric transformation, which is approximated by Lie-algebraic representation. The transformation group is mapped to a vector space which in turn feeds the statistical modeling and inference step. Their main advantage is that spatially varying fields can be estimated from local motion without continuous tracking. Mehran *et al.* [223] use an average optical flow to advect a grid of particles that is superimposed on it. The interaction force among the particles is estimated by the Social Force Model (SFM) [126] and mapped to obtain a force flow for every pixel. By selecting random spatio-temporal volumes of force flow, normal and abnormal behavior is classified. Ren *et al.* [303] introduce the concept of entropy, from the information theory discipline, to represent spatio-temporally the individual pixels and the whole image by what they call *behavior certainty*. They also use as basis flow model and particles, reporting superior results than common methods to detect abnormal behaviors in crowds. Benabbas *et al.* [31] model direction and magnitude information from flow vectors to learn the dominant orientations and magnitudes at each spatial location of the scene. At the end, they apply a region-based segmentation based on local blocks of both information to detect the major motion patterns.

One of the most simple and common region-based representation is the so-called *blob*. Collins *et al.* [68] classify moving blobs into single human, human group, vehicle, and clutter. The features used by the three-layer neural network classifier are related to blob characteristics such as area, dispersedness, and apparent aspect ratio from blob's bounding box. Lipton *et al.* [198] also use the blob metrics for classification between humans, vehicles and clutter, but include temporal consistency constraints to improve performance. Kuno *et al.* [180] use a specialization of the blob to extract shape parameters of human silhouette. Mohan *et al.* [230] combine different example-based detectors trained to find different parts of the human body, and after ensuring that these parts are present in the proper structural configuration, a second exemple-based classifier is used to detect if it is or not a person.

The analysis of motion is essential for human activity analysis since it permits to obtain spatial and temporal changes along a video. The extraction of motion can be subdivided into two methods: *motion correspondence* and *optical flow*. From both methods different representations can be formulated. Motion correspondence deals with extracting interesting points. The correspondence for multiple frames results in a trajectory, that is considered a vector valued function. Their parametrization can be done in several ways and permit to identify important invariant motion events. However, they are no suitable to represent context-based information by themselves. Optical flow computes the displacement of each pixel between frames and its representation may consider the whole image or just a region of interest. Information extracted from its derivatives

or based on information theory concepts prove to be useful to discriminate local patches, but segmentation is difficult when multiple objects are present. Both approaches present advantages and drawbacks. Their selection is content-based, namely type of scene, camera viewpoint and recording settings, and task-based, namely type of features, local or global representation and type of motion to recognize.

2.2.2 Patterns Detection and Identification

Semantic behavior detection and learning from observing activities in video is one of the most difficult challenges in surveillance scenarios [173]. In automated visual surveillance systems, the detection and identification of common activities, and the reliable discovery of suspicious or abnormal human behavior, is of undeniable importance. Both tasks involve the modeling and classification of human activities within certain rules, which is not easy due to the randomness and non-linear nature of human movement. A solution for this problem is to partition the human movement into primitive states that can be detected and classified accordingly with the specific domain of interest [46], for instance gesture recognition, trajectory classification, motion patterns segmentation, among others. Even so, such approaches may be very dependent of the techniques that are specific only for the current application domain.

Therefore, Ivanov *et al.* [141] propose to divide the problem into even smaller domain independent subproblems. Keeping in mind the outline of a surveillance system (see Fig. 2.2), the subproblems are as less application-dependent as the lower the processing level they belong to. At the high-level vision, the tasks will perform according with the concepts and relations defined by the domain of the current application. When no *a priori* information is defined, the system may be able to interpret the information acquired so far and learn hidden relationships, in an unsupervised manner, to detect common and unusual events.

Independently of the supervision level, the surveillance system needs to know how to represent and recognize behaviors corresponding to different types of concepts as defined by Bremond *et al.* [46]: i) *basic properties*: where the feature of the object of interest is defined, such as its trajectory, motion divergence factor, gaze, etc; ii) *states*: which describe the static or dynamic situation of the objects of interest; iii) *events*: that define a change of state in the temporal domain; iv) *scenarios*: are the most high-level definitions and represent a combination of states, events or sub-scenarios. Once the concepts defined, the system should be able to match temporally an unknown sequence with a group of labeled references or learned behaviors. The efficiency of the behavior matching process mostly resides in the structure of the representation, that can cope with small variations of the feature data within each class of motion pattern. The most common analytical methods for matching time-varying data in the literature are:

a) *Dynamic Time Warping* (DTW), is a template-based dynamic programming matching technique, which computes the non-linear warping function that optimally aligns two variable length time sequences as long as time ordering constraints hold [236];

b) *Hidden Markov Models* (HMM), is a stochastic state machine used for modeling generative sequences by a set of observable sequences with spatio-temporal variability [295]. HMM are

superior to DTW in processing unsegmented successive data [370];

c) *Linear Dynamical Systems* (LDS), is a continuous state-space generalization of HMMs with a Gaussian observation model [85];

d) *Finite-State Machine* (FSM), is a model of behavior composed by a finite number of states, transitions between those states, and actions;

e) *Nondeterministic-Finite-State Automaton* (NFA), is a FSM where for each pair of state and input symbols, one or more states may be possible;

f) *Belief Networks* (BN), is a graphical model that encodes complex conditional dependencies between a set of random variables, and which are represented by local conditional probability densities (CPD) [269];

g) *Neural Network* (NN), inspired by the biological nervous system, is composed by a large number of highly interconnected processing elements, *neurons*, that work tie-together [118];

h) *Self-Organizing Neural Network*, is a type of NN that is trained under unsupervised learning when the object motions are unrestricted, using a neighborhood function that preserve the topological properties of the input space [56];

i) *Agent-based*, decomposes the learning into interactions of agents with simpler behaviors and rules.

Apart of the aforementioned methods, there are some works that apply typical methods for dimensionality reduction such as Principal Component Analysis (PCA) to model and recognize atomic activities [396]. Also some variants from HMM and NN are presented such as Markov Chain Monte Carlo (MCMC) [106] and continuous HMM [27]. Other graphical models are also explored like Dynamic Belief Networks (DBN) [135], Petri Nets (PN) [108], and Probabilistic Petri Nets (PPN) [10]. More high-level approaches that express the structure of a process using a set of production rules, drawing a parallel to grammars in language modeling, such as Context-Free Grammars (CFG) [316] are also used in combination with HMMs and BNs. In terms of unsupervised learning, there are also different approaches like a data-driven Bayesian clustering [48], Latent Semantic Analysis (LSA) [266], Fuzzy C-Means [400], and several variants of clustering [233] .

As explained previously, detection of patterns encompasses the representation and the method that models and detects a normal or abnormal change of state along the temporal domain. The literature in this topic presents different ways to combine the representations described in Section 2.2.1 with the aforementioned analytical modeling and matching methods. Contrary to previous works in video scene understanding that learn location-specific models, Oh and Hoogs [254] cluster trajectories into semantic activity models independent of scene location. They incorporate scene context entities, such as main entrances of buildings and location of static objects of interest, and build a feature vector that capture the relationships and interactions of the trajectories with those entities. Superior results than standard location-dependent clustering are reported (see Fig. 2.3). Senior *et al.* [327] combine video understanding with transaction-log to detect several events such as returns fraud, cashier fraud, customer counting, and merchandising effectiveness (see Fig. 2.4).

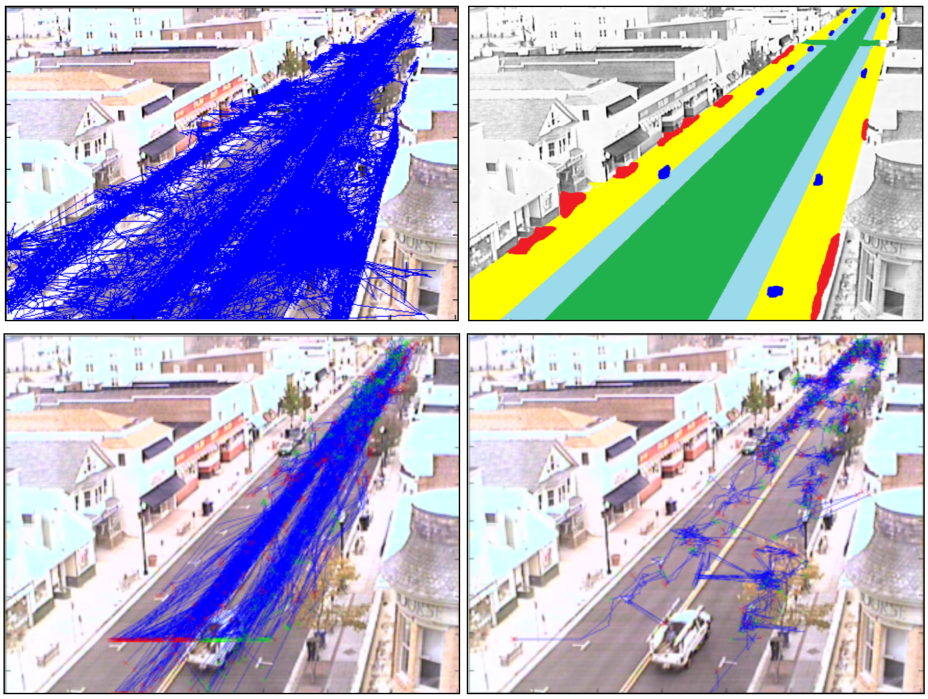


Figure 2.3: Example of an application of video scene understanding. Trajectories overlaid, scene context information as *a priori* knowledge, and two detected motion patterns: vehicles driving straight (bottom-left), vehicles parking (bottom-right). Green and red marks indicate trajectories' starting and ending points, respectively. Extracted from [254].

Figure 2.4: Example of video analytics for retail domain. Customer counting and tracking and analysis (facial expressions) for fraud detection. Extracted from [327].

Stauffer and Grimson [340] present a monitoring system that learns patterns of activity from observations of moving objects. Their work mainly focus on background subtraction and motion tracking in highway scenarios. Since the obtained representation is stable and complete, they use simple features from trajectories such as location, speed/direction, and size to create a codebook of representations based on vector quantization. Then, they accumulate the joint co-occurrence statistics over the codebook and perform a hierarchical classification over the accumulated co-occurrence data. Wang *et al.* [373] extend the previous work by introducing two novel similarity measures. Johnson and Hogg [156] learn a model of the distribution of typical trajectories through

a competitive learning NN. Morris and Trivedi [234] build a topological scene descriptor, from common trajectories, where the nodes are the point of interest, learned from a mixture of gaussians, and the edges the activity paths, learned from clustering of trajectories and represented by HMMs. Hu *et al.* [133] learn typical motion patterns in crowded scenarios from instantaneous motion flow field. They present a two-step process: creation of a directed neighborhood graph to measure the proximity of the flow vectors; hierarchical agglomerative clustering algorithm to group flow vectors into motion patterns (see Fig. 2.5).

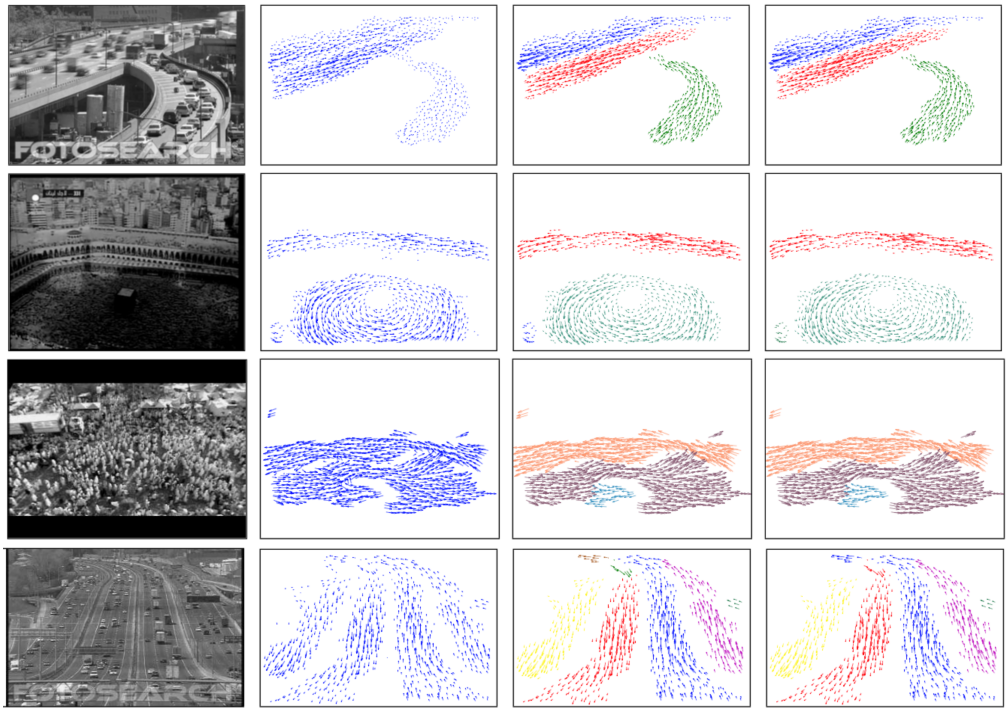


Figure 2.5: Motion patterns detected from hierarchical agglomerative clustering of motion flow field. From left to right: original sequence, motion flow field, detected motion patterns, manually ground truth. Extracted from [133].

Kratz and Nishino [175] model local spatio-temporal motion pattern in extremely crowded scenes by capturing spatial variations of local volumes in a coupled HMM, and their temporal relationship statistics are encoded in a distribution-based HMM (see Fig. 2.6). Abnormal activity is discover from statistical deviation of steady-state motion patterns. Simpler approaches such as the one of Kumar and Vaish [179], segment the scene based on velocity and direction of flow map, for further clustering of similar flow.

The approach to model and recognize a pattern should be selected according with the type of concept that the pattern represents. The different concepts defined by Bremond *et al.* [46] obey an increasing scale of complexity that involve individuals, objects, and contextual information such as interactions and scene knowledge. Modeling a complex concept, its inherent structure and associated semantic require a higher level representation and reasoning methods.

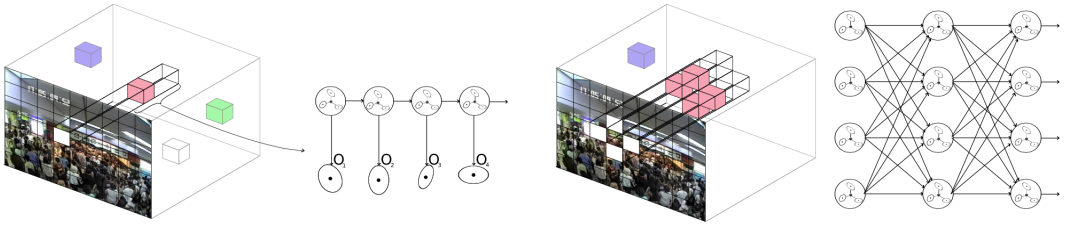


Figure 2.6: Example of a spatio-temporal HMM encoding method to detect abnormal behavior. Extracted from [175].

2.2.3 Social Behavior Analysis

Social behavior implies the interaction of two or more individuals within the same environment. The theoretical perspectives in group dynamics covers diverse branches such as motivational and emotional, behavioral, cognitive, biological, and system theory [99]. All these theories agree that a group: requires at least two people, connect people to one another, and the connection, in most cases, is socially meaningful; therefore a group is defined as *"two or more individuals who are connected by and within social relationship"* [99].

The notion of group encompasses common qualities such as *interaction*, *goals*, *interdependence of members*, *structure*, and *unity*, but the combinations of their variations redefine different type of groups, namely *primary groups*, *social groups*, *collectives*, and *categories* [99]. *Collectives* is defined as *"the aggregation of individuals that form spontaneously during a brief period of time and have permeable boundaries"*. This type of group has been tackle by the computer vision research community, namely in works associated to crowds. Their efforts have been mostly focused in technical topics like the detection of abnormal crowd behavior [175, 223, 303], group of dominant and independent motions [31, 179], statistical measurement of parameters [106, 130], detection of individuals [48], and tracking [33] (e.g. Fig. 2.7), among others. Most of these works are constrained to the requirements of the domains that normally crowd behavior analysis applies for, such as public space design, virtual environments, visual surveillance, and intelligent environments [411]. More recently, social and psychological studies have been integrated in the computer vision and artificial intelligence analysis of crowd phenomenon, starting with the support in human observations, and providing theoretical concepts that indicate new ways to represent and model people relationships in isolation and as part of a more or less dense group of people. Such theories may be categorized into *microscopic*, *mesoscopic* and *macroscopic* approaches. However, the computer vision community is still solving the problem of automatically extracting reliable and sufficient information to characterize some special crowd events.

Modeling social behaviors of people is an important branch to represent group activity. This analysis falls into the *social groups* class stated by Forsyth [99], which is defined as *"small groups of moderate duration and permeability characterized by interactions between the members over an extended period of time, sometimes goal-oriented"*. Social behavior analysis has also attracted the attention of the computer vision community, whose first efforts consist in applying *microscopic*

Figure 2.7: Example of tracking in moderately crowd scenes. Extracted from [33].

models that deal with individual pedestrians. Helbing and Molnar [126] originally introduced the SFM to investigate people movement dynamics. Ali and Shah [13] use the cellular automaton model to track in extremely crowded situations. Antonini *et al.* [18] propose a variant of Discrete Choice Model (DCM) to build a probability distribution over pedestrian positions in next time step. Pellegrini *et al.* [270] present a Linear Trajectory Avoidance (LTA) method to track multiple targets and predictions of velocities are made by the minimization of energy potentials. Scovanner and Tappen [326] model pedestrians' dynamics as a continuous optimization problem. Wu *et al.* [384] use chaotic invariants of Lagrangian Particle Trajectories (LPT) to model abnormal patterns in crowded scenes. Leal-Taixé *et al.* [189] show the importance of using social interaction models for tracking in difficult conditions. These works borrow inspiration from social and physical rules to model individual's behavior within a group, but they fail to understand the group behavior as a social entity.

More recent approaches have been crossing the border to other disciplines such as social signaling, in order to embed concepts of social-psychology into computer vision field. Bazzani *et al.* [29] consider social signal cues from face and eyes behavior, and the space component to define and detect groups. They focus in the initialization and evolution of a group, through a tracking approach that embeds the knowledge of the states of the single individuals and the state of the group. They also outline an interest map, through a visual focus of attention computed from the estimated head position and orientation. Their work is based on a simple social scenario with actors, and most of the group states are the well-known Free-Standing Conversational group (FSCG), or F-formations in the sociological terminology. Ge *et al.* [107] identify small groups of individual that are travelling together in moderate crowd scenes. They discover the groups using a bottom-up hierarchical clustering based on a symmetric hausdorff distance, which is defined by pairwise proximity and velocity parameters. Their work represents an approximation to the granularity of social groups, and their qualitative and quantitative ground-truth was validated by "human consensus" of multiple human coders, through visualizations and local interviews to pedestrians.

This validation process represents a novel approach in the literature (see Fig. 2.8). However, their approach should be tested in more demanding surveillance settings.

Figure 2.8: Example of detected small groups, marked with trajectories of different colors. Extracted from [107].

Other works investigate the addition of contextual information to improve the group activity recognition performance. Lan *et al.* [182] jointly capture the group activity, the individual person actions, and the interaction among them (see Fig. 2.9). They explore latent variables to model *group-person interaction* and *person-person interaction*, proving that these contextual information with adaptive structures optimize the inference step. Amer *et al.* [17] address a novel problem of multiscale activity recognition in a high-resolution video that allows digital zoom (in/out). They build a three-layer AND-OR graph to jointly model group activities, individual actions, and objects of interaction. The inference step process exploits the hierarchy of the graph and is cost-sensitive and scalable. Some works build local spatio-temporal descriptors from the individual person action to capture the behavior of surrounding people nearby, such as pose and motion [65, 181] (see Fig. 2.10).

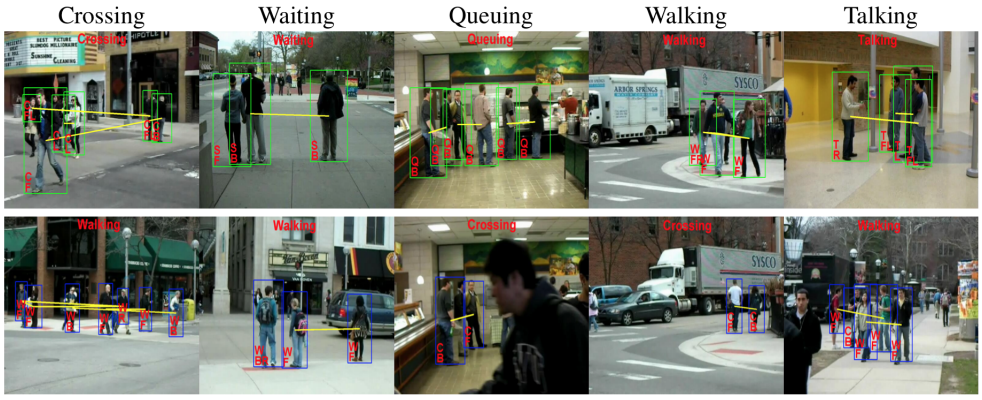


Figure 2.9: Example of detected group activity. Top row: correct classifications; Bottom row: wrong classifications. Extracted from [182].

Other approaches model the probabilistic group membership of the individuals by individual trajectories [412], while others model the underlying group structure by pairwise spatio-temporal tracking information [63]. Ni *et al.* [245] use different relations within, between and among individual trajectories to propose an encoding based on self-causality, pair-causality, and group-causality, respectively.

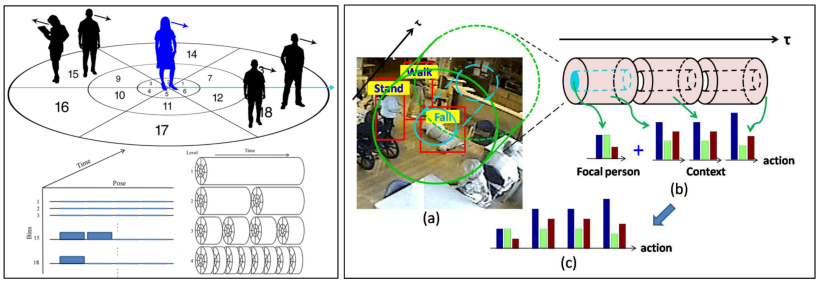


Figure 2.10: Example of context descriptors for group activity classification. Extracted from [65] (left) and [181] (right).

The identification of complex relations involved in groups and groups activities just from a video is very hard. On visual surveillance, due to camera constraints, just some features related to behavioral cues are possible to extract. Without taking into account social factors such information cannot be translated into algorithms that process and learn social meaning. Technical areas, as computer vision and artificial intelligence, and social areas, as psychology and social signaling, may collaborate for a new layer of video surveillance research.

2.3 Multimedia Video Settings

The exponential growth of the user generated online videos and the widespread of online communities for sharing video-based content such as YouTube, have been leading to an increased interest in the research community to provide solutions for the automatic detection of complex events in unconstrained multimedia videos [201, 335]. A complex event is defined as a higher level semantic abstraction than concepts such as objects, actions, or scenes, presented in a long video [394] (see Fig. 2.11). For example, an event denominated as *"attempting a board trick"* may contain multiple objects' concepts such as skateboard, surfboard, snowboard, etc; may take different scene place such as outside, skate park, etc; may de defined by related human activities like standing, jumping, sitting or laying on the board, etc; and may be accompanied by audio such as sound of board hitting the surface.

Multimedia event detection (MED) is a multimedia retrieval task with the goal of classifying each video with a particular event in an internet video archive, given the correct label of the event, example videos (positives, negatives and *related-exemplars*) and descriptions. The Text Retrieval Conference's (TREC's) Video Retrieval Evaluation (TRECVID), funded by the National Institute of Standards and Technology (NIST) and other US government agencies, has been promoting the

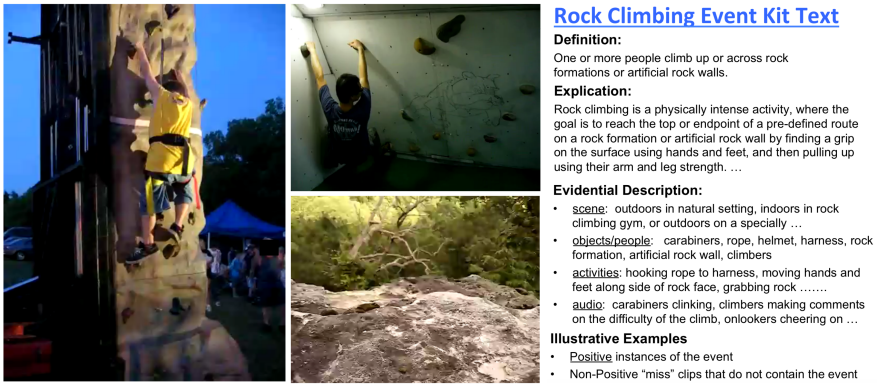


Figure 2.11: Example of MED event kit description.

video analysis and retrieval evaluation via an annual contest where organizations and individuals worldwide share their knowledge and approaches, and common metrics and reliable benchmarking about their performance are measure and make publicly available. Due to the complexity of this task, it is difficult to outline a general system diagram. As an example, Fig. 2.12 illustrates the system proposed by the CMU team for TRECVID MED 2014 challenge. However, common approaches normally accomplish such demanding task by combining multiple features from different sources, and one of the difficulties is how to effectively join those features by appropriately mining their correlations and highlight their complementary attributes [394].

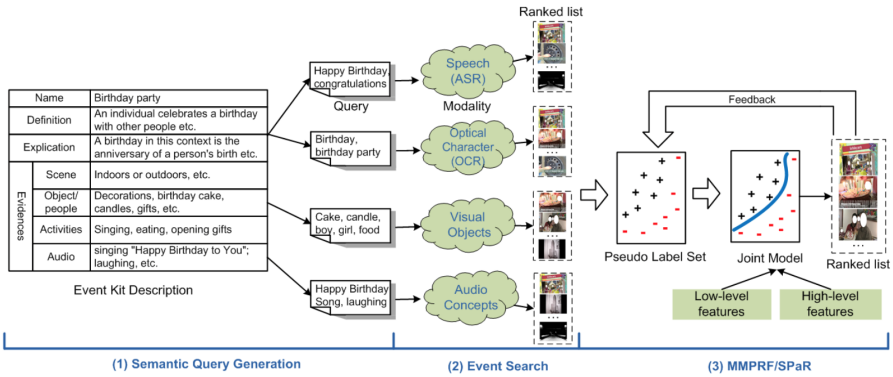


Figure 2.12: Example of the system proposed by the CMU team for TRECVID MED 2014.

The vast majority of the works in this domain focus on one of the following approaches: i) design of highly discriminative and robust features; ii) combination of multiple features from different complementary modalities such as visual, audio, and text. Next, both approaches are presented as spatio-temporal representations, feature perspective, and event detection techniques, classification perspective, respectively. Important remarks and conclusions from relevant work in this domain are also highlighted.

2.3.1 Spatio-Temporal Representations

Human action detection in complex scenes is a difficult problem due to the high-dimensionality of the search space and dynamic background. Spatio-temporal information is mandatory for this challenge. Wang *et al.* [371] represent a video as a set of feature trajectories and model human action as the spatio-temporal tube (ST-tube) of maximum mutual information. The action regions are inferred by a one-order Markov model, and the ST-tube is built by concatenating the consecutive action regions bounding the human bodies. They report superior results from their counterpart methods, spatio-temporal cuboid model [83,309], in datasets such as KTH[1] [324], CMU and UCF sports[2] [309].

Gilbert *et al.* [109] deal with the problems of camera motion, human appearance variation, scale, occlusions and background clutter, through dense corner features that are spatially and temporally grouped in a hierarchical process to produce an overcomplete compound feature set. Then, data mining is used to discover reoccurring patterns. Oikonomopoulos and Pantic [255] use spatio-temporal Local Steering Kernel (LSK) features, which capture the local structure and dynamics of the underlying activities, to build a hierarchical representation of mined dense spatio-temporal features, and at each level the formed constellations of features are increasingly discriminative of a specific action class.

The presence of specific objects can help to identify the event of interest. Deformable part models (DPM) [95] has achieved state-of-the-art performance for object detection. Niebles *et al.* [246] represent activities as temporal compositions of motion segments, which are encoded by their appearance model. However, spatio and temporal components are separately modeled and matched. Tian *et al.* [351] present a generalization of deformable part models from 2D to 3D spatio-temporal volumes. For each action model, the most discriminative 3D volumes are selected and their spatio-temporal relations are learned. This approach is able not only to classify actions, but also to localize them (see Fig. 2.13).

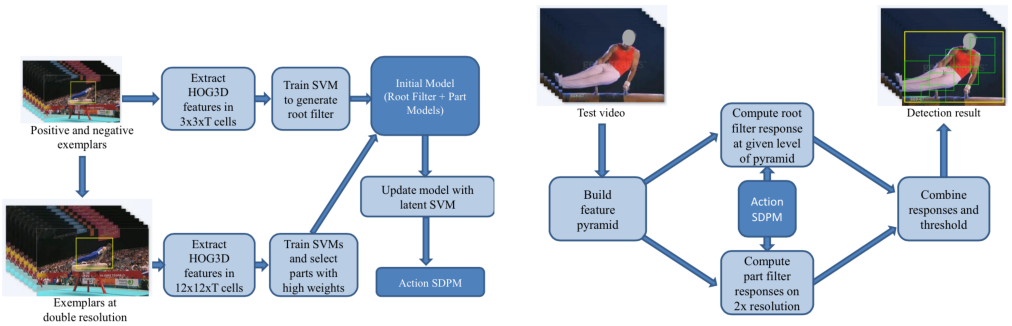


Figure 2.13: Example of the 3D spatio-temporal deformable part model. Left: training step; Right: testing step. Extracted from [351].

[1] http://www.nada.kth.se/cvap/actions/
[2] http://crcv.ucf.edu/data/UCF_Sports_Action.php

Video segmentation generalizes the concept, from image segmentation, of grouping pixels into spatio-temporal regions that exhibit coherence in both appearance and motion. This generalization presents common problems: i) *temporal coherence*: long-term coherence is mandatory to obtain reliable 3D hierarchy levels, spatial region matching is not enough to guarantee consistency over time; ii) *automatic processing*: manual selection of regions of interest and their progress over temporal domain is not known *a priori*; iii) *scalability*: the larger the video, or the temporal interval assumed, the larger the amount of pixels or features to account for, therefore the scalability of algorithms that enable efficient and reliable segmentation of long videos is a major issue.

The temporal coherence is normally obtained by using just past information or by using both past and future frames. Patti *et al.* [268] consider past information by using Kalman filters to aggregate data over time. Wang *et al.* [368] treat the video as a 3D space-time volume, therefore they consider past and future data, and use anisotropic kernel mean-shift for segmentation. Zitnick *et al.* [417] employ a tracking-based video segmentation which statistically model an image pair by using appearance and motion constraints. Interactive object segmentation approaches report high quality segmentations results, but are driven by the user input [296]. Other methods aim to obtain top-down segmentation through user input related to category-specific information. Most existing approaches to semantic video segmentation are graph-based. Jain *et al.* [143] propose an efficient coarse-to-fine energy minimization strategy that explores the spatial and temporal coherence of the initial supervoxels [392]. They use supervoxel tree hierarchy such that most supervoxels at the coarse level correspond to a single label, pruning the labeling search space for subsequent finer levels (see Fig. 2.14). Grundmann *et al.* [114] propose an automatic iterative hierarchical method that uses a graph-based approach. It is driven by dense optical-flow to guide the temporal connections. The generated segmentations are temporally coherent with stable region boundaries, and different levels of granularity are provided.

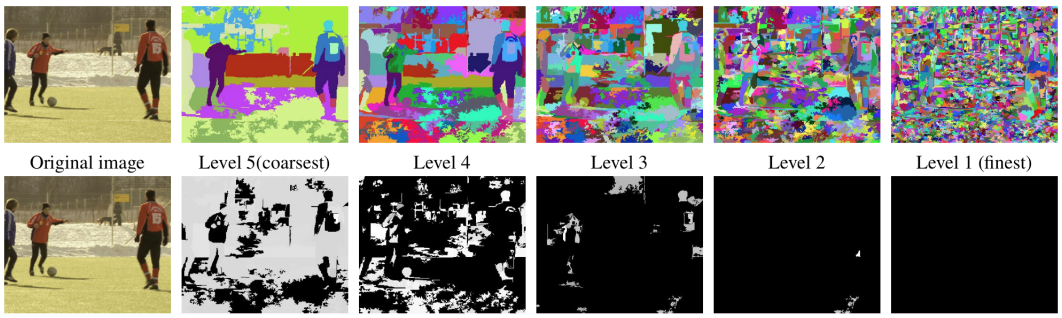


Figure 2.14: Example of the supervoxel hierarchy and portion of the search space in the supervoxel-tree explored by the coarse-to-fine scheme to find the optimal labeling. Extracted from [143].

Some approaches intend to build a global video descriptor for classification. Common methods require the localization of human body, background subtraction or tracking. The most used representations are silhouettes [40] and motion [37]. Torralba *et al.* [258] propose a global scene

descriptor based on power spectrum features, but it is not suitable for action recognition. Solmaz *et al.* [339] build their proposal over the assumption that the hybrid use of motion and static features improve classification performance of action recognition. They apply a bank of 3D spatio-temporal filters on the frequency domain of the video, integrating both information about motion and scene structure. However, they need to combine their descriptor with a local one to obtain the best classification performance.

Spatio-temporal interest point-based (STIP) methods have also received attention by the research community, since they represent the scene and the actions as a combination of 2D/3D local descriptors computed from a neighborhood around the points of interest. Klaser *et al.* [167] present a video descriptor based on histograms of oriented 3D spatio-temporal gradients. Scovanner *et al.* [325] propose a 3D-SIFT. Natarajan *et al.* [240] propose a multichannel shape-flow kernel descriptor for event detection. They extend kernel descriptors to the spatio-temporal domain to model salient flow, gradient, and texture patterns from different color channels. Such local descriptor approaches are less sensitive to noise or occlusion, but they require a large amount of relevant points and are not able to model global geometrical or temporal information.

Trajectory-based methods are also common to represent action. Gaidon *et al.* [102] propose a hierarchical action structure of mid-level motion components. They create an unordered binary tree from a divisive clustering algorithm. The input is the local trajectories and the output is the tree modeled by nested histograms of local motion features extracted along the trajectories. Matikainen *et al.* [215] quantize trajectories snippets of tracked features. The work of Wang *et al.* [364,366] is the most known and used in the literature. They extract motion trajectories, from matching point of interests through consecutive frames, and represent the video by computing Motion Boundary Histogram (MBH), Histogram of Gradients (HOG), Histogram of Optical Flow (HOF) and trajectory descriptors along the trajectories. The improved version of this work deals with camera motion compensation and the removal of trajectories that belong to background. Jiang *et al.* [148] extend the previous work of Wang *et al.* [364] to use global and local reference points to characterize motion information from local-patch trajectories, capturing motion relationships robust to camera movement. Jain *et al.* [144] also use space-time trajectories but their local descriptors are based on kinematics features of flow such as divergence, curl and shear. They decompose motion and remove the residual component to accurately extract the trajectories. However, any of these approaches capture scene or context information.

Spatio-temporal representations are crucial for the representation and modeling of complex events in challenging multimedia videos. Different approaches have been used to detect relevant motion parts and track their properties over temporal domain, while keeping temporal coherence and processing scalability. Recent works have stated that the longer the representation, the better the recognition. Semantic meaning, such as objects, may be associated to motion parts or can be separately extracted from local detectors. Relationships among the detected motion parts also play an important role in the description of the action. Normally, these approaches need to be combined with appearance features and/or global descriptors to obtain superior classification results.

2.3.2 Event Detection

Many research papers state that a multimodal approach help to improve the classification performance on video [241]. Two type of combination strategies are possible, early and late fusion. The former combines features before the classification step, such as multi-kernel learning [69], while the latter combines the output of classifiers from different features, such as average fusion, committee voting [357], among others. There is no universal conclusion on which one is the preferred strategy for multimedia content. However, Snoek *et al.* [338] conduct a large experiment and take two important conclusions: for most of the learning concepts, late fusion scheme tends to perform better, but it comes with the price of an increased learning effort; if early fusion gives better performance, the improvements are more significant. Later, Ayache *et al.* [23] verify on TRECVID 2006 data, that early fusion obtains better results on most of the concepts, while late fusion is more robust on harder concepts. To incorporate the advantages of both methods, Lan *et al.* [184] present a double fusion scheme (see Fig. 2.15). They apply early fusion to generate various combinations of features from subsets on the single features pool. Then they train classifiers on each feature combination and perform late fusion on the output of those classifiers. They also point out two important conclusions after their experiments in TRECVID MED 2010 and 2011 data: weighted is better than average combination for late fusion but not for early fusion; how to learn weight for early fusion is still an open question. Liu *et al.* [202] design a local expert forest model for score fusion from multiple classifiers under heavily imbalanced class distributions (see Fig. 2.15). Their likelihood-space is sensitive to local label distributions, and multiple pairs of locally optimized experts, on different partitions, are trained to form the forest, balancing local adaptivity and over-fitting of the model.

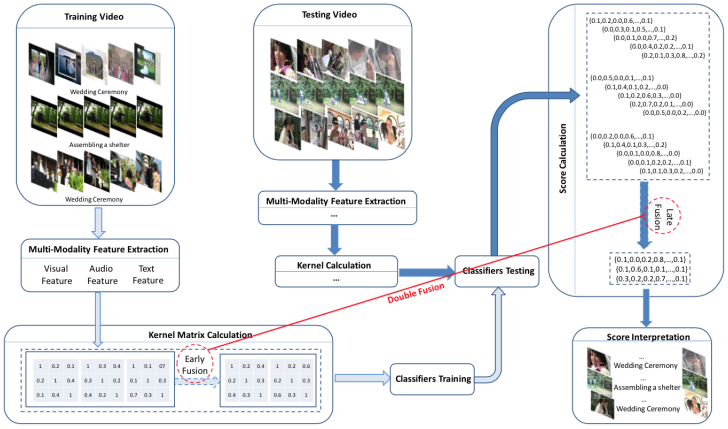


Figure 2.15: MED system with a *double fusion* scheme to improve classification results. Extracted from [184].

Fisher vectors (FV) [259] and Vector of Locally Aggregated Descriptors (VLAD) [145] have been superimposed to bag-of-words techniques for the integration of low-level descriptors. Dense local trajectories [364, 366] have been one of the most successfully representations for action recognition in the literature. One of their disadvantages is that they capture few local motion

features of important body parts in most actions. Ni *et al.* [244] work over the assumption that stressing the local features associated with important motion parts of the respective actions will lead to a better and more discriminative action representation (see Fig. 2.16). Motion parts are generated by spatio-temporal clustering of trajectories considering an action class discriminative term, their weights are learned and used in a weighted Fisher vector representation.

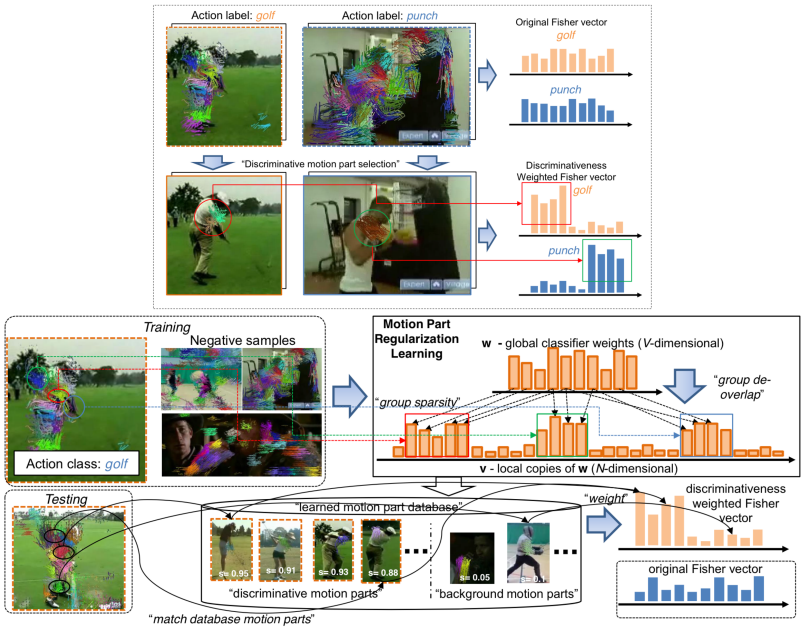


Figure 2.16: Motion part regularization to generate discriminative weighted Fisher vector representation. Illustration of the involved pipeline. Extracted from [244].

Izadinia *et al.* [142] subdivide complex events into low-level events to be treated as latent variables into a latent Support Vector Machine (SVM) model. Ma *et al.* [207] exploit transfer learning for recognition of complex events when few positive examples are available. Current approaches propose the use of deep Convolutional Neural Networks (CNNs) to advance the performance of event detection. Xu *et al.* [395] test CNNs, from where just frame level static descriptors can be extracted, and propose latent concept descriptors to generate more discriminative CNN descriptors. They also investigate the best encoding method to aggregate features. They report the best mean average precision (mAP) on TRECVID MEDTest 2013 and 2014 datasets. Gan *et al.* [103] follow the same approach and propose the so-called Deep Event Network (DevNet) that simultaneously predicts the event class and provides key spatial-temporal evidences (see Fig. 2.17). The first step is training with CNN features at the key frame level to detect the event of interest. The second step generates a spatio-temporal saliency map, which is used to find the most relevant frames of the event and corresponding location of the objects of interest.

Other works deal with a particular problem, the so-called *related-exemplars*, which are videos that share some positive elements of the current event, but have no uniform pattern due to the huge variance of relevance levels among them, therefore the relatedness assessment is subjective.

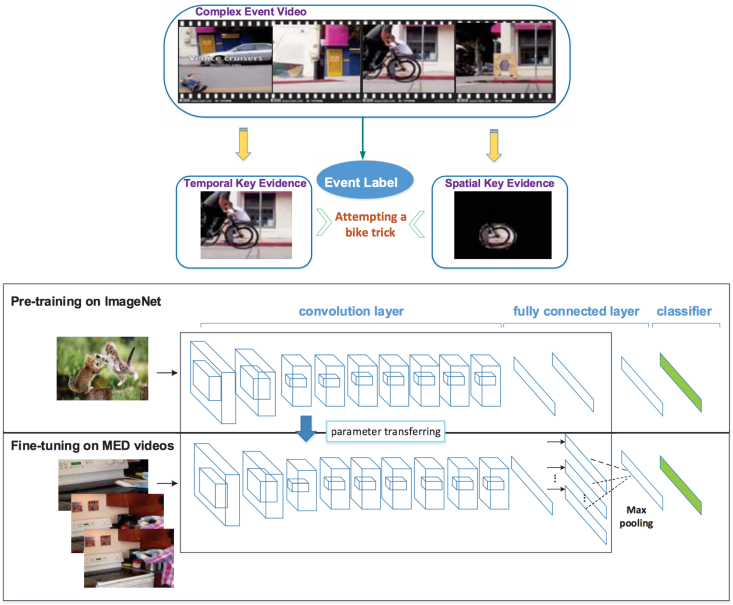


Figure 2.17: DevNet provides event label, as well as spatio-temporal key evidences. DevNet layout: pre-training step using the ImageNet; fine-tuning on the MED video dataset. Extracted from [103].

Solving this problem is beneficial when few training examples are available. Yang *et al.* [401] measure how positive the *related-exemplars* are for the current event being detected, and use them on an exemplar-specific basis. Their algorithm is designed for just one feature. Xu *et al.* [394] extend the solution to multiple features and propose a cross-feature voting scheme to explore the relevance levels of each *related-exemplar*, which are represented by ordinal labels.

The detection of complex events from multimedia videos with uncontrolled settings, such as camera motion, human editing, and background clutter, are typically associated to high-dimensional features, imbalanced labeled data, and wide intra-class and inter-class variation. The inclusion of multimodal features permits a better handling of semantic meaning across the different event classes. When a large feature set is present, the most suitable fusion technique should be able identify and ignore the non-discriminative features with respect to a particular event. Providing a single event label for a video may not be enough, therefore most recent works aim to retrieve informative temporal and spatial evidence that lead to the detection decision.

2.4 Behavioral Settings

After decades of neglect, due to the supremacy of the learning theory and then the cognitive revolution, the research on emotion, behavior and expressiveness analysis was leveraged by the important work on facial expression by Tomkins [352,353] and continued by Ekman [89,90]. The power of nonverbal behavior in emotional, or simply relational episodes, became a central issue in most psychology textbooks as these started to be invaded by photos with prototypical expressions

and simple emotions [122]. In 1992 appeared the concept of basic or fundamental emotions. It was created by theorists in the area of behavior analysis during the course of their works. Although researchers considered that human behavior can be segmented in a set of fundamental emotions there was no unanimity in which ones these are [88]. This topic remains currently a doubt among people who dedicate themselves to the analysis of human behavior.

Nowadays, the panorama regarding the nonverbal concomitants of emotional experiences has changed drastically. Every year, more than 50 books and papers are published featuring nonverbal channels of expressive communication, included in a wide range of different areas of expertise. The channels considered are mainly facial expression, gestures, gaze, vocal quality, paralinguistic features, posture and body position, head nods, among others [122]. Even though much of present-day expressive research is carried out with paper-and-pencil assessment of verbal reports of expressive content, the computer vision field is trying to emerge as a relevant tool for achieving nonverbal sensitivity in a computational and automatic way. This ability is named *nonverbal sensitivity* or *emotional intelligence*, defined as the ability to encode or express and to decode or understand nonverbal cues [321]. It would be of great value if machines were able to detect and interpret temporal patterns of nonverbal behavioral cues in several social situations like sensing agreement or disagreement among a group of people arguing about a certain topic.

The best way of achieving this goal is to successfully carry the inherent knowledge that humans have on this matter to computer systems, allowing these systems to sense activities and social relationships. However, understanding the context of a visual environment is essential to properly interpret behavior, since the context will be distinct for each application. If we think about it, this concept of context evaluation is inherent to the human condition once we have a constant necessity of adjusting ourselves to the situation we are experiencing [379]. Context encompasses spatial and temporal knowledge, but also interpretation of the functionality of the object and intention involved in the action. Computer vision research on Human Behavior Analysis (HBA) includes a broad range of studies on developing computer systems and models to achieve nonverbal sensitivity in different contexts and through different channels such as face, voice, gait and body gesture. HBA is increasingly more of interest for computer vision and artificial intelligence researchers [60]. Message production and processing, relational communication, social interaction and networks, deception and impression management, and emotional expression are the main applications for nonverbal sensitivity in computer systems [225].

The general system for human activity analysis under behavioral settings (see Fig. 2.18) normaly involves the following phases: a) scene recording; b) detection of individuals; c) extraction of multimodal behavioral cues; d) encoding and classification in terms of social signals; e) sensing the context while the scene is being recording; f) interpretation of the classified social signs considering the context constraints and classification of the output into the target social-behaviors.

Following is presented a review about some of the most important socio-psychological conclusions for understanding body behavior in the literature, highlighting the importance of gestures, posture, environment and their interpretations. Some technical approaches that aim to detect social-behaviors are also briefly described.

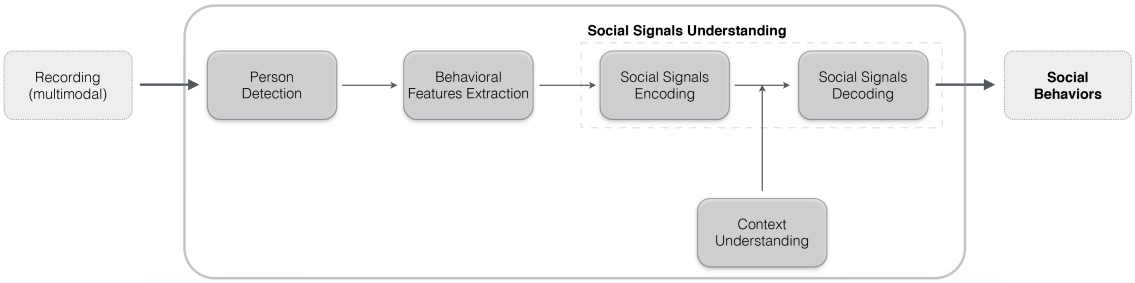


Figure 2.18: General system for social signals and behavior analysis.

2.4.1 Understanding Body Behavior

A behavior representation needs to include both cumulative and temporal information about the object of interest. A behavior model should benefit more from temporal information about behavior, in contrast to an object model that mostly incorporates spatial information [110]. The intuition behind is to capture a significant influence of the correct interpretation of what the individual is about to do [8]. During any human interaction, there are many relations between human body parts and the surrounding environment. The key for understanding behavior is then to analyse human interactions based on well defined existent relationships [311].

Body expression is an inherent feature of human behavior. Many artists use it in exaggerated ways to express their ideas to the public in a nonverbal way. If they are expressive enough, the message they are trying to pass on is received by that public. Whole-body expressions provide information about the emotional state of the producer, but also signal his action intentions. For example, a surprised body expression can signal the appearance of a new element in a scene but also give information on how the subject will deal with that surprise.

The work in body expression was initiated in 1872 by Darwin who described in detail the body expressions of many different emotions [76]. Recent psychology studies have pursued Darwin's preliminary work. De Gelder *et al.* [78] describe the stimulus set of whole body expressions termed Bodily Expressive Action Stimulus Test (BEAST), providing validation data through the creation of a database composed of 254 whole body expressions from 46 actors expressing 4 emotions. Van den Stock and Righart [82] investigate whole-body expressions of emotions in three different experiments from which realized the importance of emotional whole-body expressions in communication, either when viewed on their own or in combination with facial expressions and emotional voices. Kleinsmith *et al.* [169] propose a method that automatically recognizes affective states and affective dimensions from non-acted body postures and use observers to establish ground truth labels. The review paper by Kleinsmith and Bianchi-Berthouze [168] provides an overview, gathering and evaluation of the main state-of-the-art studies on body expressions as a communication channel.

Studies in psychology indicate that the combination of facial expression and body gesture are more informative than each alone [221]. Body gestures have been indicated in the literature [218]

as illustrators, emblems and regulators social signs, since they are associated as much as 90% with speech. In some cases, they regulate interactions, communicate a specific meaning, or stress a discourse [155]. Gesture expressiveness can be also transmitted unconsciously [124], for instance head inclination, shifting posture, and face touching are often attached to social affective states such as shame, discomfort and embarrassment [91], while the manipulation of small objects can be interpreted as self-protection [171].

Postures affirm the current attitude of people towards social situations. Scheflen [322] propose three criteria to determine the classification of postural behavior: i) *inclusive* vs *non-inclusive*, measures how much a given posture considers the presence of others; ii) *face-to-face* vs *parallel-body-orientation*, measures the degree of engagement in a conversation; iii) *congruence* vs *incongruence*, measures the level of psychological involvement (i.e. affiliation or rejection). Depending on the conversational layout, physical distances between individuals affect their relationship, and, consequently, influence their body expressiveness. Hall [120] dictates the interpersonal distances around the individual as concentric zones, which measure the level of intimacy and are defined by the self-explanatory terms: *intimate*, *casual-personal*, *socio-consultive*, and *public*. The environment influences social interactions and, consequently, behavior. Russo [315] exposes interesting characteristics about the connotation of seating arrangements, e.g. extrovert people tend to keep shorter interpersonal distances than introvert ones.

From a technical point of view, the first experiments on vision-based gesture recognition were conducted on sign-languages applications [386]. Later, De Silva *et al.* [81] recognize children's emotion in the context of a game by an affective gesture recognition system. Piana *et al.* [287] present a preparatory work that automatically performs tracking and analysis of nonverbal expressive cues to assist children with Autism Spectrum Conditions (ASC) (see Fig. 2.19). Since they use Microsoft Kinect, they are able to detect and track points from the skeleton of the individuals. They group the points and extract low-level motion features such as velocity, acceleration, bounding volume and kinetic energy. Mid-level features are composed from those features and represent concepts such as repetitivity, impulsivity, smoothness, contraction, and symmetry. How to infer higher concepts remain an open-question in their work.

Gesture recognition measures the similarities between hand motion pattern characteristics (see Fig. 2.20). Some works assume that each type of gesture is unique, therefore they perform the comparison by matching trajectory templates [40]. However, most gestures are not well-defined in real conditions, their starting and ending points are different from person to person, they may suffer interruptions from occlusion or interactions (non-linear temporal scale), and different gestures may exhibit similar motion patterns (ambiguity in temporal segmentation). Psarrou *et al.* [298] recognize human gestures and behavior by a statistical framework that learn the prior, using HMM, and continuous propagation of density models of behavior patterns. Burgoon *et al.* [53] identify emotional states from bodily features, and De Silva and Bianchi-Berthouze [80] recognize emotions from statistical representations of posture features.

There is an evidence that nonverbal cues can be correlated with human intent, however the situational body language can be difficult to interpret [52]. In close-range situations, facial expres-

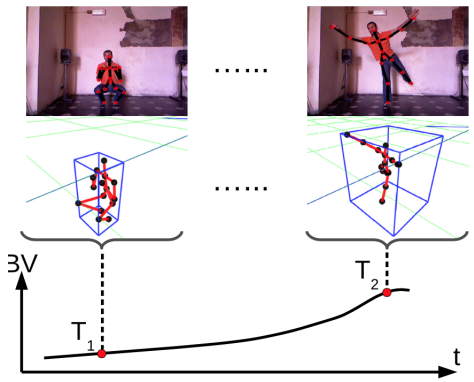


Figure 2.19: Example of the bounding volume surrounding an individual, from which low-level and mid-level features are extracted. Extracted from [287].

sions and eye movement patterns are the main cues for correlating intent, while in far-field distance, body orientation and head pose are the most informative measurements. Many efforts have been done to estimate body and head pose in low-resolution images [32], crowd scenarios [260], and by estimating articulated body parts [96]. Human gait can also help to identify intent, since it characterize periodic motions by spatio-temporal patterns of human silhouettes [28].

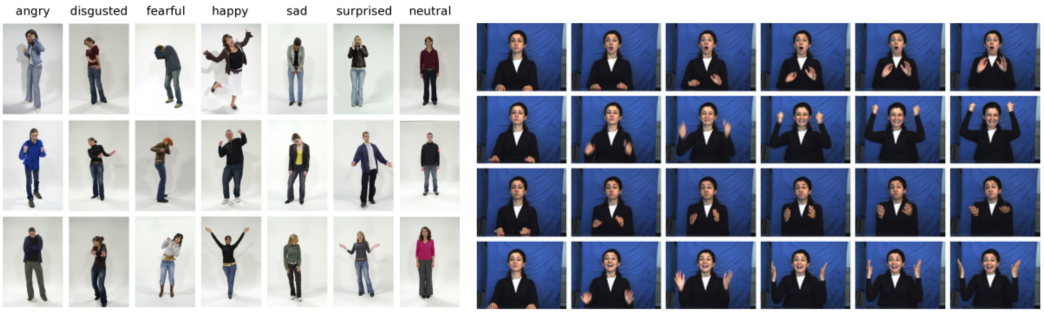


Figure 2.20: Example of body language and body gestures (FABO dataset [116]) from top to bottom: fear, joy, uncertainty, and surprise.

Observation and automatic interpretation of behavior is a complex process due to both heterogeneity and dynamicity of data. Visual analysis of nonverbal cues such as head pose, body orientation, posture and gait, all play an important role to interpret intent of a person and to describe nonverbal human behavior. Models of nonverbal behaviors in social interaction environments are useful to develop assistive technologies to support the deficits of specific populations and bring them into fuller participation in the general community.

2.5 Discussion

Automatic recognition of human activities is an important area of research in computer vision and artificial intelligence with applications in many diverse fields like surveillance, HCI, biometrics,

multimedia retrieval, assistive technologies, and others. The interest for this field started in the early 1970s with the study of motion as main drive force of human action. Nearly two decades after, human activity recognition systems begin to be included in the visual surveillance industry field with some experimental prototypes delivered by important researchers in the field, such as Prof. Dr. J.K. Aggarwal[1] and Prof. Dr. Mubarak Shah[2]. Nowadays, technology related to this field is being tested on multiple specific areas such as home-healthcare, Advanced Driver Assistance Systems (ADAS), forensic technologies, among others.

Three main steps in understanding human activity are clearly identified: *extraction* of motion information, *representation* for encoding and discriminative reasons, and *interpretation* for recognition, inference and forecasting purposes. The research community has been tackling with low-level problems since a long time ago, and progress has been made in the representation of information. Recent works have been introducing the use of context to improve high-level knowledge of interpretation, therefore, in understanding better the human activity.

Representation plays a crucial role on visual intelligent systems. It describes an abstract entity, highlighting and encoding its main characteristics. Some relevant researchers [329], under the vision understanding field, sustain that motion is the main source of information for a reliable interpretation of human activity, much more important than structural information recovery from structure-from-motion approaches. In general, the representation should be as invariant, compact, and robust as possible for further classification or recognition task, where, normally, the model is compared, based on a distance metric, with an unknown input. The problem resides on the fact that the system is just able to recognize a predefined set of behaviors, and cannot learn new behaviors from them. The solution passes through an incremental system that learns primitive models to be able to infer more complex models of activities from observations.

Some of the proposed systems assume non-real conditions and others adopt straightforward computation processes, which is a sign that this field is still looking for improvements. Since the complexity of this field is very high, due to the wide low-level-to-high-level processing chain involved, initial solutions accept several subproblems as solved, providing manual annotation data, to keep the high-level goal treatable. Special efforts should be done by the research community to integrate small processing modules into the overall system to automate the complete processing chain and evaluate the impact of each module in the final output. Real-case scenarios are also mandatory to provide accurate and reliable performance reports. Also, despite of recent advances on training methods applied on different areas, the classification of human activity still does not have an universal solution. Most of this is due to the inherent intra-class variation of human motion, the same happens with activities of identical semantics that also have large intra-class variation, the diversity of scenarios with large environmental variations, and the multiple recordings settings.

Human activity recognition is the foundation for human behavior understanding, which requires additional, and multimodal, contextual information for the W5+ concepts (who, where,

[1] http://cvrc.ece.utexas.edu/aggarwaljk/index.html
[2] http://crcv.ucf.edu/people/faculty/shah.php

what, why, and how) [264, 265]. Indeed, the same activity may have different behavior interpretations depending on the social rules and context in which it is performed. A natural step toward social, nonverbal and contextual cues is being pursued in this field in collaboration with other disciplines. The main challenges in HBA are related to its complexity and uncertainty, and developing learning strategies for HBA requires a deep knowledge in psychology of emotions in order to build behavior models and patterns for extrapolation to several types of analyses.

There are many technical problem to solve, specially challenging for real-world deployment. Having access to uncompressed, non-interlaced, high-resolution, and high-frame-rate video can be beneficial not only to improve current solutions, but also to open new research challenges [329]. It is also necessary to use multiple-cameras to track people for an extended period of time, since the field-of-view of a singular camera is very limited. However, this stands two important issues to deal with: networking communication and superimposed field of view among the cameras. Indeed, tracking is one of the common processing steps of this type of systems. Much effort have been done in this topic, but it remains difficult to accomplish a reliable solution for all the possible type of scenarios. In some cases, if this tracking information is not correctly extracted, it will impair the recognition of the activities being performed. The change of human action appearance increase the difficulty to detect human body parts and decrease the classification accuracy of atomic actions. From an industrial point of view, is mandatory that object segmentation, tracking and detection can be reliably computed with ordinary devices. How to enable better and more reliable ways to include context information in an automatic way is crucial not only to social-aware applications, but also to detect behaviors in individual and collective activities in surveillance scenarios and to add semantic meaning to motion relevant parts in the classification of videos.

Many upcoming trends can be outlined due to the importance of this field in the current society, for instance: tracking people in crowds in complete occlusion with shadows and variable lighting; recognize human activities from moving cameras; combination of information with multiple non-visual sensors; learn models that cope with varying body shapes, clothing, occlusion and clutter in real-world conditions; how to effectively model jointly facial and body behavior; deploy intelligent systems capable of self-adaptation to sudden changes and to recover from failure; embed real-time implementations in specific hardware support for the hot-trend products related to augmented reality and virtual reality. However, the future direction of research will be dictated by the future requirements of the applications. Public and private spaces are becoming more and more ubiquitous and sensitive, which has been giving rise to new applications that include social signaling and affective/intentional understanding.

Chapter 3

Long-range Motion Trajectories*

> The longer you look at an object,
> the more abstract it becomes, and,
> ironically, the more real.
>
> Lucian Freud

Nowadays, almost any public space has a Closed Circuit Television (CCTV) system installed. This has fostered the implementation of (semi-)automatic systems to interpret real-world scenes by monitoring pedestrians and their activities, detecting common motion patterns, and identifying unusual behaviors. The key insight for this type of system is to exploit spatio-temporal relationships among entities and motion patterns, while retaining structural information of the scene. Underlying motion representations such as trajectories are, by nature, intuitive and useful to provide solutions for such problems.

3.1 Introduction

Motion is a fundamental cue for scene analysis and human activity understanding in videos. Considering the surveillance settings, motion can be encoded in trajectory structures useful for tracking objects in complex scenarios and for action recognition, or it can be used in the form of flow to address behavior analysis in crowded scenes. Normally, each approach can only be applied on limited scenarios, lacking a more generic solution in the literature.

There are various types of scenes in the surveillance settings, from dense crowd context to multi-tracking with sparse groups, whose distinction is defined by the *scene density* and the *object size*. Following a trajectory-based approach, long-duration trajectories offer several advantages [345] over short-range tracklets [215, 301, 363] for visual analysis tasks such as activity discovery and learning of semantic region models for event recognition. However, they imply to overcome occlusions, camera motions, and nonrigid deformations issues. To avoid ambiguity with related

*Some portions of this Chapter appeared in [277–279]

work, the term **global motion trajectories** is referenced here as the trajectories that translate typical paths of pedestrians, normally used for video scene analysis [156, 372].

This Chapter presents our analysis of motion in surveillance settings to suppress the difficulty in finding a generic and robust approach that can provide a useful spatio-temporal representation to the analysis of motion patterns in different and highly variable surveillance scenarios. It is important to emphasize that, in this application, the goal is not to characterize the motion pattern for each pedestrian; instead we aim at characterizing the prototypical motion patterns of the pedestrians within the scene. The challenges derive not only from the recording and viewpoint conditions, but mainly, from the content type of the scene, which varies due to the number of pedestrians, randomness of their movements, scene cluttering and scene layout. Considering our categorization defined in Chapter 1, this study lies on *the group* level providing a global representation of human motion in surveillance settings.

Our proposal is a generic motion-based framework, so-called **VIdeo-based LOn-range Motion Analysis** (***VILOMA***), that represents the spatial and temporal features of the flow in terms of long-range global trajectories. We highlight the contributions, namely the framework formulation, the generic approach to handle scene variability and motion context variations, and its capability to integrate motion from local and global representations. The emphasis of its outcome is represented by the long-range motion trajectories that overcome trajectory-based approach problems, as well as the diverse framework characteristics that can be used for human activity tasks, with special focus on motion pattern segmentation.

3.2 Overview

3.2.1 Microscopic and Macroscopic Approaches

Human motion can be described at different scales. Each one implies specific motion analysis since their underlying relationships among pedestrians and space-context behavior differ. Normally, for high density scenes and low object resolution, motion is modeled at a global level and patterns are inferred [11, 133, 222]. On the other side, for scenes with a small number of objects, multi-tracking approaches [175, 413] are preferred since they track objects individually and describe motion by their spatial position. The computer vision community has been addressing several research problems related to each scenario independently.

Crowded scenes present two types of categories, *structured* and *unstructured*, depending if the movement of objects are defined by physical constraints or if they move freely in any direction, respectively. Related work focuses on modelling scene structures and on recognising the co-occurrences of crowd behaviors. Ali and Mubarak [11] propose a framework that implements a Lagrangian Particle Dynamics (LPD) to advect a grid of particles and use them for motion interpretation in the form of physically and dynamically distinguishable motion segments. This type of approaches overcomes the lack of optical flow in capturing long-range temporal dependencies,

and do not suffer from problems faced by object-tracking-based approaches. However, they do not consider spatial changes, cause time delays, and imply high computational effort.

For *structured* scenes, motion patterns are the most salient features that help understanding the scene [413]. Some approaches [164, 175] consider the division of the video in spatio-temporal cuboids to identify prototypical motion pattern representations and variations within each one. It is not usual to extract motion trajectories from this type of conditions. However, to the best of our knowledge, there are some works [13, 132, 263] that try to approximate the extraction of motion patterns in terms of *super tracks*, but they did not measure their similarity with traditional object tracks and they did not test their approach on low density scenes.

For *unstructured* scenes, the concept of *coherent motion* emerges. It describes the free collective movement of pedestrians in groups and try to infer collective behaviors. This type of approach models the crowd dynamics focusing on pedestrians and interactions among them [414]. These approaches can follow two types of taxonomy: i) *macroscopic* studies [415], that consider groups as a collective and homogeneous block where the individual is transformed by the group; ii) *microscopic* approaches [223], which analyse groups as the composition of individual agents that interact with each other and with the environment. Macro models have statistical meaning and not physical, while micro present physical validation but are difficult to scale up to macro scale.

Scenarios with sparse and dense groups follow a single or multi-tracking approaches. Both present difficulties related to target's size, number of similar objects, and occlusions [413]. For crowded scenes, tracking-based models disregard the correlation between pedestrians in a close vicinity. Motion trajectory mechanisms can also be performed at feature level [215, 363], instead of object level, by tracking interest points. However, they face critical factors to solve: selection of good tracking features, correct mapping between selected features and actions of interest, trajectory discontinuity due to inconsistent point correspondence, among others.

3.2.2 Trajectory Analysis

Motion information can be encoded in trajectory structures useful for tracking objects in complex scenarios, as well as for action recognition. Sun *et al.* [346] extract trajectories by matching Scale-Invariant Feature Transform (SIFT) descriptors between consecutive frames, and propose a three-level hierarchical framework to describe the spatio-temporal trajectory-based context: point-level, intra-trajectory, and inter-trajectory context. Messing *et al.* [224] represent trajectories by velocity history of tracked Harris3D keypoints, compare their features with Dollar *et al.*'s spatio-temporal cuboids [84], and Laptev *et al.*'s space-time interest points [185] to show their effectiveness with the state-of-the-art work. They prove that velocity history feature can be extended and combined with other useful information. They use Birchield's implementation [36] of the Kanade-Lucas-Tomasi (KLT) tracker, remove the affine consistency check to maximize the tracking duration of feature points and allow non-rigid motion. Matikainen *et al.* [215] consider quantized trajectories snippets of tracked features obtained from a KLT tracker. Two variants are explored: one that it is a simple concatenated vector of derivatives, and other that combines the previous vector with a

vector of local affine transformations. They argue that this approach avoid the pitfalls of existing dense optical flow based trajectories.

Lezama *et al.* [190] show that long-term motion analysis brings important cues for higher-level scene understanding in terms of objects and event categories. Their novelty is the inclusion of the depth ordering factor that treats the occlusion and disocclusion relations among tracks, on the track clustering cost function. Their work is based on Brox and Malik's research [50], which presents consistent spatio-temporal segmentations of moving objects through the definition of pair-wise distances between long term point trajectories. Multi-body factorization methods [70,92,300,399] analyze trajectories of 3D rigid objects over time by exploiting the properties of an affine camera model, but suffer of non-Gaussian noise and on partial occlusion and disocclusion motion.

Semantic interpretation can be build using trajectory information as input. Pusiol *et al.* [299] design an intermediate layer composed of Primitive Events descriptors, which represent motion occurring between topology slow regions. A topological region of interest is obtained by clustering individual segments of slow points in trajectories. At the end, meaningful transitions between topological regions are captured. Long-term observations of moving objects in the scene allow to segment them into semantic regions and build semantic scene models from the spatial distribution of trajectories. Automatic learning of the geometric and statistical models of structures in the scene can be obtained by clustering trajectories based on spatial and velocity distributions [374].

Trajectories represent the motion history of well-defined points of interests. Their parametrization may encode useful information not only in terms of motion, but also in terms of relationships among other points of interest and/or scene entities. The most difficult issue to solve is the correspondence problem between consecutive frames, specially under occlusions, noise, clutter background, and camera motion.

3.2.3 Optical Flow

Horn and Schunck [129] distinguish motion field and optical flow concepts. Their taxonomy defines motion field as the 2D projection of the 3D motion of objects in world, and optical flow as the apparent motion of brightness patterns in the image. Optical flow algorithms formulate the problem as the optimization of a global energy function represented by

$$E_{Global} = E_{Data} + \lambda E_{Prior} \tag{3.1}$$

where the data term (E_{Data}) measures the consistency of the optical flow within the sequence of images, and the prior term (E_{Prior}) works as a regularizer for the Aperture Problem [26].

The baseline of data term can be based on the brightness constancy assumption or on the optical flow constraint equation. Normally, the optical flow constraint is obtained from the linearization of brightness constancy during the optimization step. The brightness constancy assumes that

the pixel's color intensity does not change when it flows from one image to another, and it is expressed as

$$I(x,y,t) = I(x+u, y+v, t+1) \tag{3.2}$$

The linearised version of brightness constancy assumption yields the optical flow constraint as

$$I_x u + I_y v + I_t = 0, \tag{3.3}$$

where subscripts refer to partial derivatives. Both equations include one constraint on the two unknowns at each pixel and introduce one error per pixel [26], so a penalty function is associated. The baseline is the L_2 norm that assumes Gaussian and independent and identically distributed (IID) errors for the optical flow constraint. Extensions of the data term consider photometrically invariant features such as gradient information [49], descriptor features [267], explicit illumination modeling [226], and multi-band information [320].

Since the data term is ill-posed, the prior term favors certain flow fields over others. The simplest smoothness prior considers small gradients of the flow field. The computation of this term also involves the choice of a penalty function that normally lies into Total Variations (TV) methods [49], that avoids false assumptions postulated by L_2 norm. Some refinements consider the inclusion of weights in the penalty function with a spatially varying function that relate flow discontinuity with position on the region [343], for instance, flow discontinuity is higher on edges that inside the region. Normally, the weighting function is isotropic, but automatic anisotropic weighting is another approach [416]. Further approaches replace the first order derivatives baseline with higher-order priors [250].

The regularization of motion field during flow estimation is a difficult problem, which is closely related to aperture phenomenon. Indeed, the main problem of optical flow computation is how to design an anisotropic smoothness regularizer that maintains both variable spatial coherence into a region, and accurate flow discontinuities at the motion boundaries [390]. Wrong flow estimation arises from common regularizer operator's problems dealing with non-rigid scenes, where objects may have irregular deformation, occlusion regions, and poor texture scenes.

All the recent optical flow algorithms have a common formulation: an objective function that combines a data term, which assumes constancy image properties, and a spatial term, that models flow motion. Sun *et al.* [344] present a detailed performance comparison and analysis of the objective function and the optimization method of the most accurate methods on the Middlebury flow dataset[1]. They investigate the reasons behind differences on behavior performance of existent optical flow algorithms and conclude that: simple derivative constancy is enough for pre-filtering the image; a graduated non-convexity (GNC) scheme is useful; spline-based bicubic interpolation presents great consistency; a non-convex Charbonnier penalty function is preferable; median filtering between each warping iteraction to denoise intermediate flow field improve accuracy. The underlying of their work can be used to develop new optical flow algorithms.

[1] http://vision.middlebury.edu/flow/

The analysis and evaluation conducted by Baker *et al.* [26] present important conclusions over accuracy and performance of optical flow algorithms: refinements on both terms of global energy function, such as spatial weighting, illumination modeling, and usage of features, improve accuracy; variational optimization approaches perform better than gradient-descent; algorithms that use rigidity perform poorly on non-rigid scenes; continuous optimization techniques are preferred than discrete optimization due to search efficiency for fidelity in sampling process. They conclude that the big problem for optical flow algorithms are the large motion discontinuities and fast motion of complex objects.

3.2.4 Flow Dynamics

Motion can be described by Lagrangian and Eulerian flow descriptions, which are formulated on different frames of reference and describe coherent structures of temporal dynamics in terms of trajectories. The Lagrangian coordinate system implies the advection and tracking of particles injected into the flow, and permits the observation of how the flow deforms and rotates the fluid. The Eulerian approach extracts a dense flow coverage since particles are computed at fixed positions, providing an overview over the entire flow at a specific time instance.

In a time dependent vector field there are four types of characteristic curves: streamlines, pathlines, streaklines, and timelines. Streamlines and pathlines are described as curves tangent to the vector field. Streaklines can be computed from the spatial and temporal gradients of the flow map. For unsteady flows, directions of flow depends on time as well as on position therefore streamline, pathline, and streakline representations are different [152]. In this work, we explore the streaklines and streamlines complementary representations.

3.2.5 Streaklines and Streamlines

Streaklines are the locus of points that connect all the particles that had been originated from the same initial point in the past at a given time. Streaklines should not get too long due to shape inconsistency with the flow and instability on numerical integration solution. Mehran *et al.* [222] reveal that streaklines are the most informative flow representation when compared with optical flow and particle flow. *Streak Flow* can be obtained from time integration of the velocity field. Such representation fills the gaps of optical flow and captures faster immediate dynamic flow changes than traditional particle flow representation [222].

Streamlines can be obtained by bidirectional numerical integration of the vector field using an autonomous ODE (Ordinary Differential Equation) system. They can be described as curves tangent to the vector field at every point in the flow [376]. The integration starts from a seed point and ends when it: reaches another streamline's neighbour or a critical point, hits the domain boundary, or forms a closed path. Several streamline placement algorithms have been proposed in the literature including flow topology based methods [358], evenly-spaced streamline placement method [153], and a hybrid flow topology-evenly-spaced streamline algorithm [383]. All of them share three common stages: i) seed placement, ii) diffusion process, iii) stopping criteria.

3.2.6 Vector Field Representation and Advection

A grid of particles is overlaid on the flow field. The scene's motion is quantified by particles' movement driven by dense optical flow. This advection process considers a video represented by a 3-dimensional array $W \times H \times T$, where T is the number of frames, W frame's width, and H frame's height, and an optical flow map $(u_w(t), v_h(t))$, where $w \in [1, W]$, $h \in [1, H]$, and $t \in [1, T-1]$. The particle position $(x_w(t), y_h(t))$ at grid point (w, h) at time $t+1$ is achieved by solving

$$\begin{aligned} x_w(t+1) &= x_w(t) + u(x_w(t), y_h(t), t) \\ y_h(t+1) &= y_h(t) + v(x_w(t), y_h(t), t) \end{aligned} \tag{3.4}$$

The repetition of this process at each frame yields a family of curves that represents the particle trajectory set. Since human motion creates unsteady flow, each point can be represented by a set of pathlines, streaklines, and streamlines.

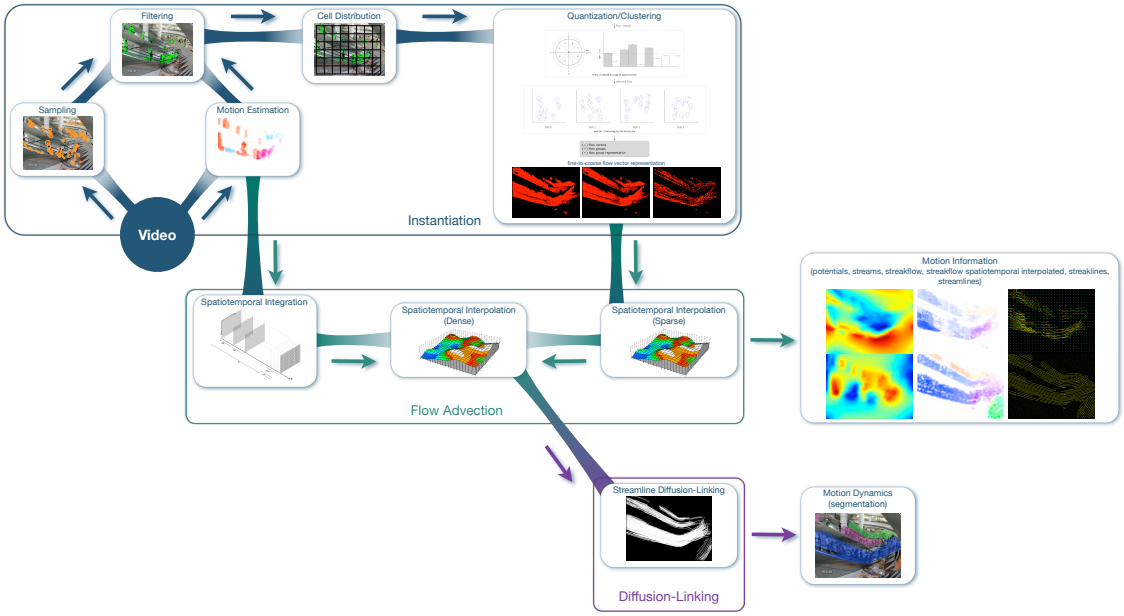


Figure 3.1: ***VILOMA*** framework for long-range motion analysis.

3.3 Motion Analysis Framework

Traditional approaches for motion analysis that consist in detecting moving objects or features, tracking them, and analyzing their tracks, miss the estimation of long-term motion representations that bring important cues for scene understanding [190]. Our approach is able to extract long-range motion trajectories that encode spatial and temporal changes in the scene, as well as local motion statistics around each trajectory point in the form of discrete distributions. It follows a Lagrangian perspective to integrate motion through the temporal domain under an Eulerian view, similar to

the *extended particle* technique defined by Mehran *et al.* [222]. The workflow of the proposed framework, so-called **VIdeo-based LOn-range Motion Analysis** (***VILOMA***), is illustrated in Fig. 3.1.

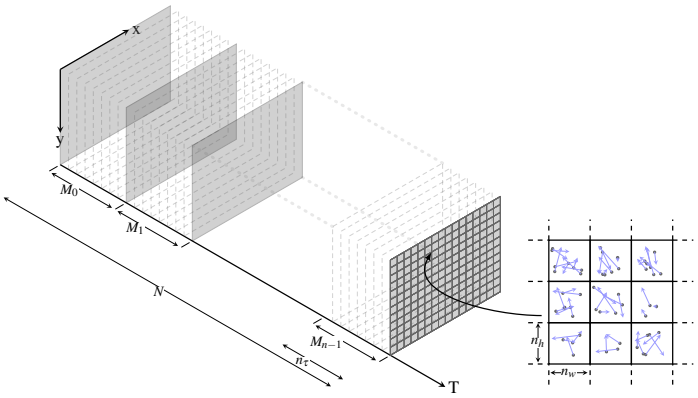


Figure 3.2: Spatio-temporal video volume representation.

The input of ***VILOMA*** is a monocular video sequence I of T frames, where its volume is $W \times H$ (pixels) $\times$ T (frames). The volume is subdivided into N spatio-temporal cells of size $n_w \times n_h \times n_\tau$, without overlap, where n_w is the width of the cell in pixels, n_h is the height of the cell in pixels, and n_τ the number of frames in the cell, which is equal to the temporal duration of each mini-batch (Fig. 3.2). The video is composed by a set of mini-batches, $\mathcal{M} = \{M_n\}$. The output primitive is a trajectory $\Gamma = (\mathbf{S}(n), \mathbf{x}(n))$, where $\mathbf{S}(n)$ is the trajectory descriptor, and $\mathbf{x}(n) = (x(n), y(n))$ are the trajectory spatial coordinates at mini-batch n. The temporal coordinate, n, is an integer (correspond to mini-batches) and the spatial coordinates, $\mathbf{x}(n)$, are in sub-pixel accuracy. The set of detected trajectories is denoted by $\Im = \{\Gamma_i\}$. ***VILOMA*** comprises several steps that are summarized in the Algorithm 1 and explained next.

3.3.1 Instantiation

The beginning of this stage is composed by the sampling strategy and by the motion estimation represented by flow maps, which are indeed instantaneous velocities. Both information are combined to create the flow vectors. After, they undergo a filtering step to remove noise and outliers, and then they are distributed among the enclosing cells where are locally quantized and grouped. These steps form the *instantiation* block of ***VILOMA*** and are executed at each frame, excepting the quantization and clustering operation. Its output is: i) an averaged flow map from the instantaneous flow maps of the current mini-batch; ii) a global fine-to-coarse flow vector representation (see Fig. 3.3). All the operations taken in this stage are executed within each mini-batch.

3.3.1.1 Sampling

This step extracts a set of key points for matching between previous and current frames. The sampling strategy could follow one of two possible distributions: dense or sparse. For image classification [192, 251] and action recognition [367] dense sampling performs better than sparse

Algorithm 1: Main Loop algorithm

```
1  Procedure Main()
2      for f = 1 to T do
           // Computes sparse/dense sampling and returns key points.
3          keyPoints ⟵ Sampling(f)
           // Computes instantaneous flow maps from Optical Flow algorithm.
4          instFlowMap ⟵ MotionEstimation(f)
           // Filtering step with outlier removal technique.
5          flowVectors ⟵ Filter(keyPoints, instFlowMap)
           // Computes an averaged flow map for the current mini-batch.
6          avgFlowMap ⟵ ComputeAvgFlowMap(f)
           // Distributes flow vectors through the enclosed cell.
7          DistributeSpatially(flowVectors)
8          if newMinibatch then
               // Quantization/Clustering for fine-to-coarse flow vector
                  representation.
9              flowRep ⟵ QuantizeAndClusterCells(cells)
               // Computes motion advection.
10             ComputeMotionAdvection(avgFlowMapPrev, avgFlowMapCurr)
               // Computes an averaged streak map for the set of mini-batches.
11             avgStreakMap ⟵ ComputeAvgStreakMap(f)
12             if nMinibatches = cellMemory then
                   // Extract sparse interpolated flow map from flow vector
                      representation.
13                 interpFlowMap ⟵ ComputeInterpFlowMap(flowRep)
                   // Extract dense interpolated streak flow from combination
                      between averaged streak flow and fine-to-coarse flow vector
                      representation.
14                 interpStreakFlowMap ⟵ SpatioTemporalInterp(avgStreakMap, flowRep)
                   // Compute streamlines from diffusion technique.
15                 streamlines ⟵ ComputeStreamlines(interpStreakFlowMap)
                   // Link broken streamlines to form representative long motion
                      trajectories.
16                 trajectories ⟵ LinkGloballyStreamlines(streamlines)
17             end
18         end
19     end
```

interest point detectors. Dense sampling enables local reasoning from motion similarities, and introduces spatial regularity constraints in the clustering method [50].

For dense sampling, we divide the 2D image space in a static grid spaced by $W = (w_x, w_y)$ pixels. Considering the aperture problem, not all parts of the image contain complete and meaningful motion information. Corners are points with high spatial frequency, but normally do not fit well on affine motion model. For this reason, we consider the good-features-to-track criterion [332]. In this way, a sampling point on the grid is valid if the smaller eigenvalue of its autocorrelation matrix is above a threshold that represents the noise criterion, such as

$$min(\lambda_1, \lambda_2) > \lambda_t, \tag{3.5}$$

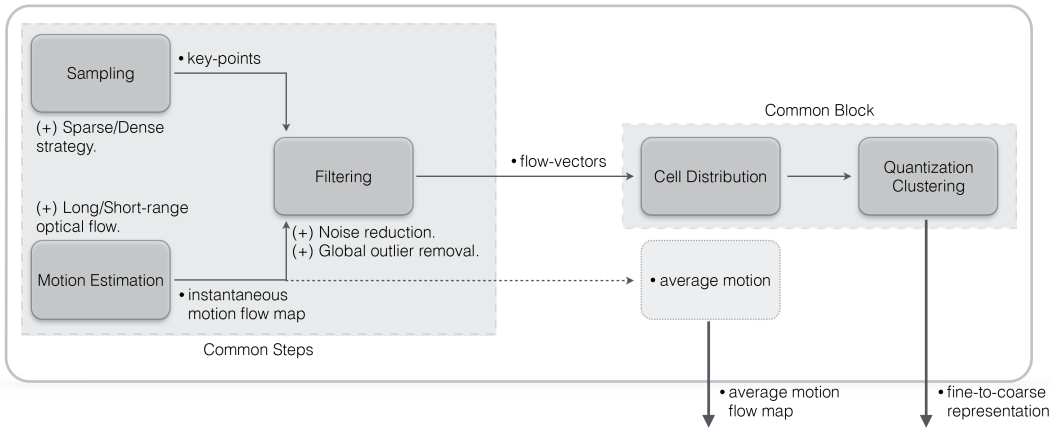


Figure 3.3: ***VILOMA*** - Instantiation step.

where λ_1 and λ_2 are the eigenvalues, and λ_t the noise threshold.

For sparse sampling, we consider a 2D space domain to avoid the joint spatio-temporal domain, since time and space have specific and different characteristics [363]. We adopt the Features from Accelerated Segment Test (FAST) sampling algorithm [310].

The dense sampling is highly computational demanding, and such effort is propagated through the subsequent steps. We also verify that it introduces noisy points that do not add discriminative value, therefore ***VILOMA*** preferably uses the sparse sampling distribution.

3.3.1.2 Motion Estimation

The motion flow is estimated using optical flow algorithms, which differ from either frame-to-frame analysis or larger spatio-temporal displacements. This step is computed independently and the resulting motion flow map is used in the filtering step process, in conjunction with the key points.

Under classical short motion displacement, we consider the following methods: i) *Pyramidal Lucas-Kanade* [406]; ii) *Farnebäck* [93]; iii) *Classic weighted non-local term* [344]; iv) *Classic equally weighted non-local term* [344]. Regrading the large motion displacement, we use the following methods: i) *Classical variational* [49]; ii) *Descriptor matching in variational model (LDOF)* [51]; iii) *Particle video* [320]; iv) *SIFT flow* [199].

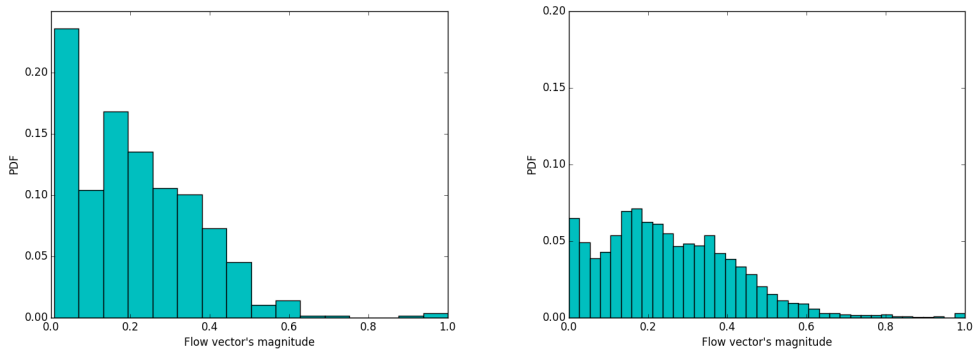
From the above algorithms, we highlight two of them: the short-classical *Farnebäck*, and the large-descriptor matching in variational model, *LDOF*. Both of them present good results: the former approximates neighborhood of two consecutive frames by quadratic polynomials, and estimates displacement fields from the polynomial expansion coefficients. Its real-time computation is an advantage, but introduces noise on areas with large appearance variations; The latter is a coarse-to-fine warping strategy that includes descriptors (Histogram of Gradients (HOG) and Geometric Blur (GB)) into the variational optical flow model to avoid local minima and formulate a numerical scheme to improve reliability for large displacements. The intuition is to combine correct large displacement correspondences from descriptor matching with variational model that

efficiently and accurately compute dense motion fields. It is an offline method that permits the extraction of smoother and longer trajectories. We adopt the Farnebäck's method due to the trade-off between computational effort and robustness.

3.3.1.3 Filtering

This step consist in the tracking of each key point from frame t to the next frame $t+1$. The two key points form a flow vector at a specific pixel location. Each flow vector is represented by $F_i = (x_i, y_i, u_i, v_i)$, where (x_i, y_i) is the key point, and (u_i, v_i) are the motion field components in x and y directions, respectively. This step builds each vector flow by assuming the key point location to be the initial vector's position, $P_i(t) = (x_i(t), y_i(t))$, and considering the flow vector's endpoint, $P_f(t) = (x_f(t), y_f(t))$, as the median (component-wise) of the flow field, $f = (u(t), v(t))$, in a neighborhood of size K. Therefore, the number of flow vectors is equal to the number of key points.

A dual-threshold on flow magnitude is applied to remove flow vectors that have little motion information, as well as extremely high magnitudes. However, we empirically verify that such operation is not enough to remove the flow vectors resulting from the background noise. Since the magnitude of the flow vectors depend on the instantaneous velocities and on the kernel size, we expect that most of the flow vector's magnitude to be smaller than the mean magnitude. Therefore, the distribution will not be symmetric and will be, normally, skewed to the right, as illustrated by Fig. 3.4.



(a) (b)

Figure 3.4: Flow vector's magnitude distribution for the Crowd Unstructured (C.U.) scenario, considering: (a) just one frame; (b) several frames.

For this reason, a novel outlier removal technique is proposed. It consists on a rough approximation that assumes a unimodal Gaussian distribution and estimates both lower and upper bounds inspired on the Chebyshev's theorem and on the skewness measure. The Chebyshev's theorem can be applied to any data set regardless of its distribution. In this case, due to the sharp skewed distribution, the Chebyshev's inequality is not good enough to estimate the non-symmetric, lower and upper, bounds, $\{\ell_-, \ell_+\}$. Our technique estimates two parameters to obtain those bounds, a

shift factor, χ, and a scale factor, ρ. The intuition behind is that the nonparametric skew measure gives a shift factor to consider the difference between the median and the mean, accordingly with the distribution's tendency (positive or negative skew), while the ratio between the number of observations represented by the median and by the mode provides a factor to scale the shift amount (the greater the ratio, the smaller the scale factor, and, consequently, the larger will be the acceptance gap between the threshold limits).

The nonparametric skew measure is given by $\gamma = (\mu - \nu)/\sigma$, which assumes values between $[-1,1]$, and where μ is the mean, ν the median, and σ the standard deviation. Considering these boundaries, the shift factor is decomposed in the following limits

$$(\chi_-, \chi_+) = \begin{cases} (\ \gamma\ , 0), & \text{if } \gamma > 0 \\ (0, |\gamma|), & \text{if } \gamma < 0 \end{cases} \tag{3.6}$$

The scale factor, $\rho = 1 - \tilde{x}/\hat{x}$, results from the relation between the number of observations represented by the median, $\tilde{x}$, and the mode, $\hat{x}$, of the log-transformed flow vector's magnitude distribution. We employ the log-transformation to approximate the data to a symmetric distribution before measuring the amplitude relation. We adopt the Freedman-Diaconis rule to obtain the optimal bin width, $Bin_w = 2 \cdot IQR \cdot n^{-\frac{1}{3}}$, where IQR is the interquartile range, and n is the number of observations in the distribution. The limits of the scale factor are given by $\{s_-, s_+\} = \{\rho \cdot \chi_-, \rho \cdot \chi_+\}, \in [0,1]$, which produce the final non-symmetric bounds, $\{\ell_-, \ell_+\} = \{\sigma \cdot s_-, \sigma \cdot s_+\}$, and that are used to estimate the final threshold limits for outlier removal, $\{\lambda_-, \lambda_+\} = \{min + \ell_-, max - \ell_+\}$, where *min* and *max* are the minimum and maximum values of the flow vector's magnitude distribution, respectively.

After several experiments on different datasets, this technique presents coherent and better results to remove global outliers on flow vector's data than state-of-the-art methods (see Section 3.4.3).

3.3.1.4 Cell Distribution

The video volume has a regular spatio-temporal distribution. Spatially each frame is divided by a grid, whose resolution is dependent on the frame size and is set at the beginning. Temporally the video duration is evenly divided. Each spatio-temporal region is denominated a cell, C_i, and contains the flow vectors whose initial positions lay inside it. Each flow vector is encoded by $F_i = (x_i, y_i, L_i, \theta_i, t_i)$, where (x_i, y_i) is the sampling point, L_i is the flow magnitude length, θ_i is the flow angle relative to the positive *x-axis*, and t_i is the frame. This step as well as the previous ones are executed every frame.

3.3.1.5 Quantization and Clustering

In order to obtain a fine-to-coarse flow vector representation that could permit to model different levels of patterns, a two-step quantization and clustering approach is applied on each cell at the end of each mini-batch. This operation considers all the flow vectors collected along the duration of the mini-batch, therefore the number of key points is much greater than the number of cells. The aim is to reduce the number of flow vectors, while maintaining the geometric structure of the flow field, and to obtain different representations of local dominant motion flows in a fine-to-coarse scale. This step is illustrated in Fig. 3.5.

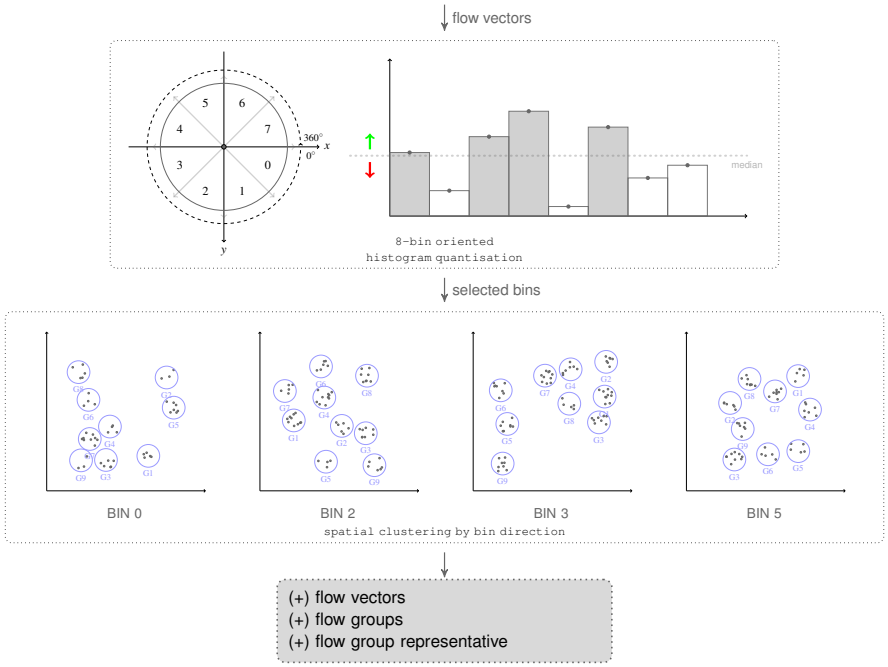


Figure 3.5: Quantization and Clustering step by cell.

The first quantization step uses the flow vector angle and considers a full-degree histogram (360°) with eight bins to represent orientation groups. Only the major groups, with weight above the histogram's median value, are taken into account for the next clustering step. This eliminates noisy flow vectors whose orientation fall apart the expected local distribution. The second step uses the flow vector position and applies a spatial clustering on each valid orientation group. A k-means approach with center initialization [20] is adopted. To select a robust k value, we compute k-means with increasing k until the compactness measure, C_k, satisfies the condition

$$\frac{C_{k+1} - C_k}{C_{k+1}} < t_c \tag{3.7}$$

with $C_k = \sum_{i=1}^{n} \|s_i - c_{l_i}\|^2$, where s_i is the input sample i, c_{l_i} is the l index clustering center to which sample i belongs, and t_c is the compactness ratio threshold, which is normally selected very low (≈ 0.01), and we set $k_{opt} = k + 1$.

Several clusters per orientation group are obtained, the so-called *dominant groups*, which are weighted by the number of flow vectors that belong to them and are ordered in a descendent-way. These groups represent the local dominant flows, which are described by $L_i = (x_i, y_i, n_i, \theta_i)$, where (x_i, y_i) is the average position, n_i is the total number of flow vectors, and θ_i is the average orientation angle. A fine-to-coarse representation, with three levels of flow vector granularity, is obtained: i) *flow vectors*, the set of all flow vectors collected and filtered, $\mathcal{F} = \{F_i\}$; ii) *flow groups*, the prototypes of the set of local dominant groups, $\mathcal{L} = \{L_i\}$; iii) *flow group representative*, the principal local dominant group, L_{rep}. They are useful for computational consumption requirements and to investigate their length and time scale impact on the dynamics of the motion advection step. This three-level global flow vector representation is obtained per mini-batch. However, it can be accumulated during the last b mini-batches, so-called *memory cell*, in order to create dense flow field representations for different discriminative levels. In this work, these representations are explored for streamline diffusion.

At this stage, each cell is able to estimate local region parameters using information from its neighborhood. One of them is the entropy, whose computation might not be reliable if a small number of samples is presented. To overcome this problem, for each cell the samples are collected, which are: its flow vectors, the *flow groups* of the cells that belongs to its neighborhood of size $K \times K$, and replicas of its flow vectors at the boundary cells. The entropy calculation relies on the probability of each angular bin x_i, $p(x_i) = C(x_i)/\sum_{i=1}^{n} C(x_i)$, where $C(x_i)$ corresponds to the number of vectors in bin x_i. Entropy is then measured and used on this work as an indicative of: i) degree of vector variation in its local neighborhood, which measures the saliency value and highlights the possibility of being a critical point or belonging to a separation line in the global vector field; ii) degree of vector variation between subsequent mini-batches, which measures the difference in the information content between both vector fields, considering the distribution shape difference and highlighting a possible new local motion pattern.

3.3.2 Flow Model Advection

This stage captures long-range temporal dependencies to represent spatial and temporal features of the flow. It is responsible to advect the motion considering an average flow map along the entire mini-batch and a dense grid of particles. It uses the global fine-to-coarse flow vector representation in a two-fold way: i) interpolated to produce sparse vector maps; ii) in combination with a sampling strategy to enrich the representation. This stage is executed at each mini-batch and its output is a set of motion representations such as streaklines, potentials, streak flow, and interpolated flow vector maps, among others (see Fig. 3.6).

3.3.2.1 Motion Advection and Spatio-temporal Interpolation

A dense grid of particles is considered. Each particle has fluid properties and their initial position correspond to each pixel on image. This characteristic follows the assumption that the computation of the streakline vector field needs a dense path line integration [376]. All particles are integrated

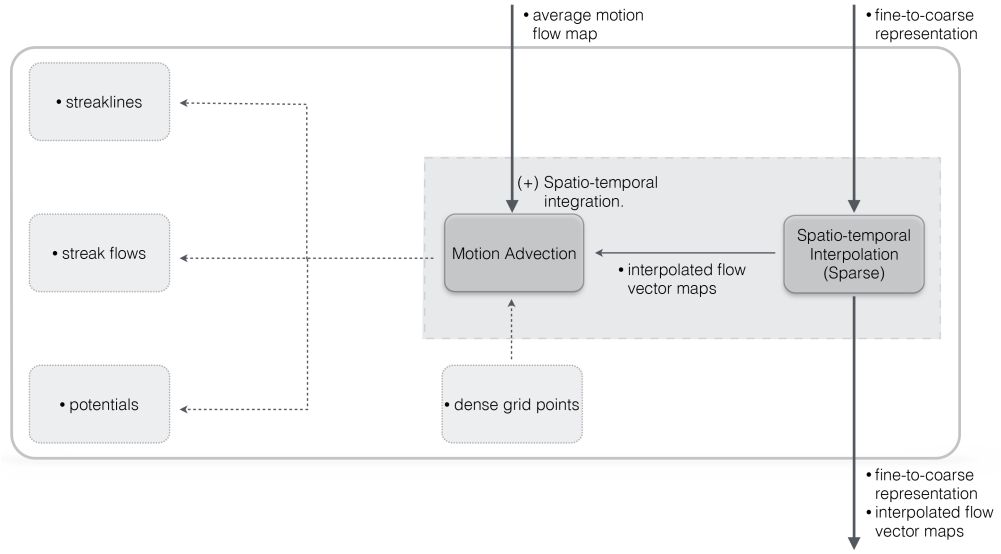


Figure 3.6: ***VILOMA*** - Flow Model Advection step.

over time accordingly to an average of the optical flow maps along the current mini-batch, which is restarted at the beginning of the next mini-batch. On each time step a particle on position p is created and all other particles previously initialized on the same position follows the flow field. This process is expressed by Eq. (3.4) and is repeated along the *memory cell* size to obtain the streaklines. The streak flow is extracted from temporal integration of the velocity field. We use the Runge-Kutta-Fehlberg (a.k.a. RKF45) for this purpose.

Considering that each streakline is a collection of particles, where each particle i is composed by a position and velocity components, we get a set of 3D data points for x and y flow directions. To compute the streak flow, $\Omega_s = (u_s(x,y), v_s(x,y))$, with sub-pixel level accuracy, we adopt a multi-resolution method based on B-spline refinement to approximate scattered data by error minimization on both dimensions. The 3D data points are given as input, the tensor product B-spline surfaces are produced, and the result is a least square approximation to the scattered data with B-splines for each flow direction that represents the streak flow on each direction (x,y) [1]. The same procedure is used to obtain the spatio-temporal interpolated flow maps of each fine-to-coarse representation, considering instead the accumulated flow vectors of each one.

3.3.3 Streamline Diffusion-Linking

Due to the ending conditions of the streamline diffusion process, described in Section 3.2.5, and since we are interest in extracting long-range streamlines to represent global motion trajectories, we include a post-processing step that links short streamlines. This stage uses the flow information collected over a sequence of mini-batches, of *memory cell* size (see Fig. 3.7).

[1] We use the Least Squares Approximation of Scattered Data with B-splines Library (http://www.sintef.no/Projectweb/Geometry-Toolkits/LSMG/).

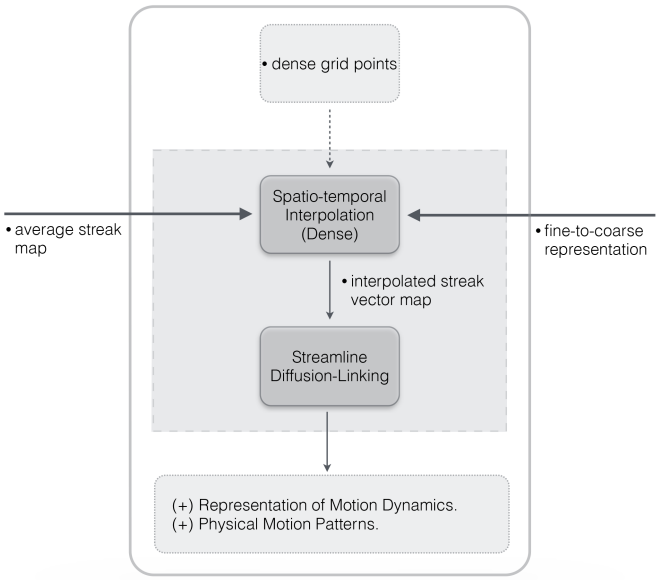


Figure 3.7: ***VILOMA*** - Streamline Diffusion-Linking step.

Streamline diffusion requires as input a vector field obtained from a dense flow field and the largest it represents the temporal and local changes over time, the longer the streamlines, and the better they emphasize the global field temporal coherency and the better they describe the topology of the flow. To this end, we use a combination of: i) the averaged streak flow from the current set of mini-batches; ii) the set of flow vectors from a specific fine-to-coarse representation collected along the current set of mini-batches; iii) a dense grid of particles. The resulting flow field is formed by B-spline interpolation, as explained in Section 3.3.2.1, considering as input the set of flow vectors from the fine-to-coarse representation superimposed to the averaged streak flow. This flow field is converted into a vector field using a grid-based discretisation and the filtering step explained on Section 3.3.1.3. We adopt the state-of-the-art farthest point seeding method [220] as the streamline diffusion technique [2].

The streamline linking process permits to obtain long-range streamlines. It is formulated as a combinatorial matching problem that considers compatibility in terms of flow appearance, motion, and spatio-temporal regularization among all short-streamlines. We adopt a discrete Markov Random Field (MRF) process to encode association constraints between *query* and *candidate* streamlines, re-correlate the set of short-streamlines and extract an optimal linkage between them. This undirected model is inspired in the work of Rubinstein *et al.* [313]. Our formulation in terms of probability of linkage, $\mathcal{L}$, between streamlines is defined by

$$P(\mathcal{L}) = \prod_{i} \phi_i(l_i) \prod_{i,j \in \mathcal{N}(i)} \psi_{i,j}(l_i, l_j), \tag{3.8}$$

where $\phi_i(l_i)$ are the unary potentials that model the compatibility between a query streamline, q_i,

[2] Algorithm is implemented on the Computational Geometry Algorithms library (CGAL) http://www.cgal.org/.

and a candidate streamline, c_j; $\psi_{i,j}(l_i, l_j)$ are the pairwise potentials for link regularisation, in case of tracking ambiguities, between a pair of query streamlines, q_i and q_j, considering the candidate streamlines that lay in their spatio-temporal neighborhood, $\mathcal{N}(i)$. The global optimization problem, given by Eq. 3.8, is inferred using a tree-reweighted belief propagation. Under this context, the streamlines taken on the MRF process are called tracks.

The compatibility term is divided into three components: a) the *appearance similarity*, ϕ_a, which models the flow properties; b) the *motion similarity*, ϕ_m, which takes into account the velocity information; c) the *prior on motion model* (aka *motion discontinuity similarity*), ϕ_p, which approximates a motion model to predict next streamline's position in case of large discontinuities. Appearance and motion similarity terms consider a symmetrically weighted average comparison of features along the last n elements of the query track, q_i, and the first n elements of the candidate track, c_j. The weight is an exponentially decaying factor, $w_t(k) = \alpha^k, 0 < \alpha < 1$, that works as a confidence parameter.

Instead of considering an individual average information (motion or appearance) for each track and then take their difference in the similarity term, as used by Rubinstein *et al.* [313], we adopt point-to-point operations. We formulate the appearance term between tracks Γ_i and Γ_j based on the cosine similarity of the streak flow's angle at each track's position given by

$$\mathbf{s}_{ij} = \frac{1}{Z}\sum_{k=0}^{n_a-1}(S_i(t_i^{end}-k)w_o(t_i^{end}-k) - \quad S_j(t_j^{start}+k)w_o(t_j^{start}+k))w_t(k) \tag{3.9}$$

where $S_i(t)$ is the track's, Γ_i, cosine of the streak flow angle at time t, $w_o(t)$ is an outlier weight that measures how well $S_i(t)$ fits the appearance characteristics of the entire track, Γ_i, which is modelled by a Gaussian distribution of the track's streak flow angles. The same is defined for track Γ_j. The normalization factor is expressed by

$$Z = \sum_{k=0}^{n_a-1}(w_o(t_i^{end}-k) - w_o(t_j^{start}+k))w_t(k) \tag{3.10}$$

and the *appearance similarity* is defined by

$$\phi_a = \exp(-\frac{1}{\sigma_a^2}\|\mathbf{s}_{ij}\|) \tag{3.11}$$

The motion term considers the velocity variation. Similarly, the velocity difference is taken by a point-to-point track relation stated by

$$\mathbf{v}_{ij} = \sum_{k=0}^{n_v-1}(v_i(t_i^{end}-k) - v_j(t_j^{start}+k))w_t(k), \tag{3.12}$$

where $v_i(t)$ is the track's, Γ_i, velocity at time t. The same is defined for track Γ_j. The *motion similarity* is expressed by

$$\phi_m = \exp(-\frac{1}{\sigma_m^2}\|\mathbf{v}_{ij}\|) \tag{3.13}$$

The *prior on motion model* that predicts track's movement on discontinuities considers linear kinematic equations to estimate the closest point of the query track, Γ_i, to the initial point of the candidate track, Γ_j. The motion integration is done until the distance travelled equals the length between the last point of Γ_i and the first point of Γ_j, and is governed by

$$\begin{aligned} \mathbf{x}_i(t+1) &= \mathbf{x}_i(t) + \mathbf{v}_i(t) + \frac{1}{2}\mathbf{a}_i(t) + \mathbf{v}_i^{flow}(t) \\ \mathbf{a}_i(t+1) &= \mathbf{v}_i(t+1) - \mathbf{v}_i(t) \end{aligned} \tag{3.14}$$

where $\mathbf{v}_i^{flow}(t)$ is the flow vector velocity at position $\mathbf{x}_i(t)$. The next velocity, $\mathbf{v}_i(t+1)$, is randomly chosen from a Gaussian distribution of the velocities of the track Γ_i. After this, a weighted distance, that includes spatial and angular values between the last predicted point of the query track and the initial point of the candidate track, is used to obtain the *motion discontinuity similarity* term

$$\phi_p = \exp(-\frac{1}{\sigma_p^2}\left(\alpha\|\mathbf{x}_j^{start}(t) - \mathbf{x}_i^{end}(t)\| + (1-\alpha)\angle(\mathbf{l}_i,\mathbf{l}_j)\right)) \tag{3.15}$$

where $\mathbf{l}_i$ is the last segment of the query track, $\mathbf{l}_j$ is the first segment of the candidate track, $\angle(\mathbf{l}_i,\mathbf{l}_j)$ is the angle between both segments, and α is a weighted factor (in this case 0.5). For each set of mini-batches, ***VILOMA*** extracts a set of streamlines, which are connected with the streamlines obtained from the subsequent set of mini-batches.

Every node in the graph, i.e. every streamline, has an additional state with a predefined cost to represent the terminal state and to avoid a forced linking. The graph just defines unary potentials among streamlines that present a *compatibility* term, ϕ_i, below a predetermined threshold. Candidate track's formation is evaluated under geometrical constraints: i) *length* (d_{thr}), which defines the spatial distance between the last point of the query track and the first point of the candidate track; ii) *continuity direction* (θ_{dir}), which returns the angle between the last segment of the query track and the segment that links the last point of the query track with the first point of the candidate track; iii) *direction difference* (δ_{dif}), which states the angular difference between the last segment of the query track and the first segment of the candidate track. Only the ones that satisfy pre-defined thresholds are included in the graph. In the same way, the *compatibility* term between query tracks, $\psi_{i,j}$, follows a geometrical pruning with the same constraints, excluding the δ_{dif} constraint. In terms of temporal neighboring, only the tracks which belong to the same set of mini-batches are considered in the link regularization step. This pruning process effectively reduces the computational effort without affecting the final results.

We took further advantage of the ***VILOMA***'s characteristics and use the cell's entropy to detect the areas with high entropy values. The query tracks whose ending points and candidate track

whose starting points fall on these regions are not considered in the linking process. This highly improves efficiency, while keeping accuracy. The overall outline of ***VILOMA*** is described in Fig. 3.8

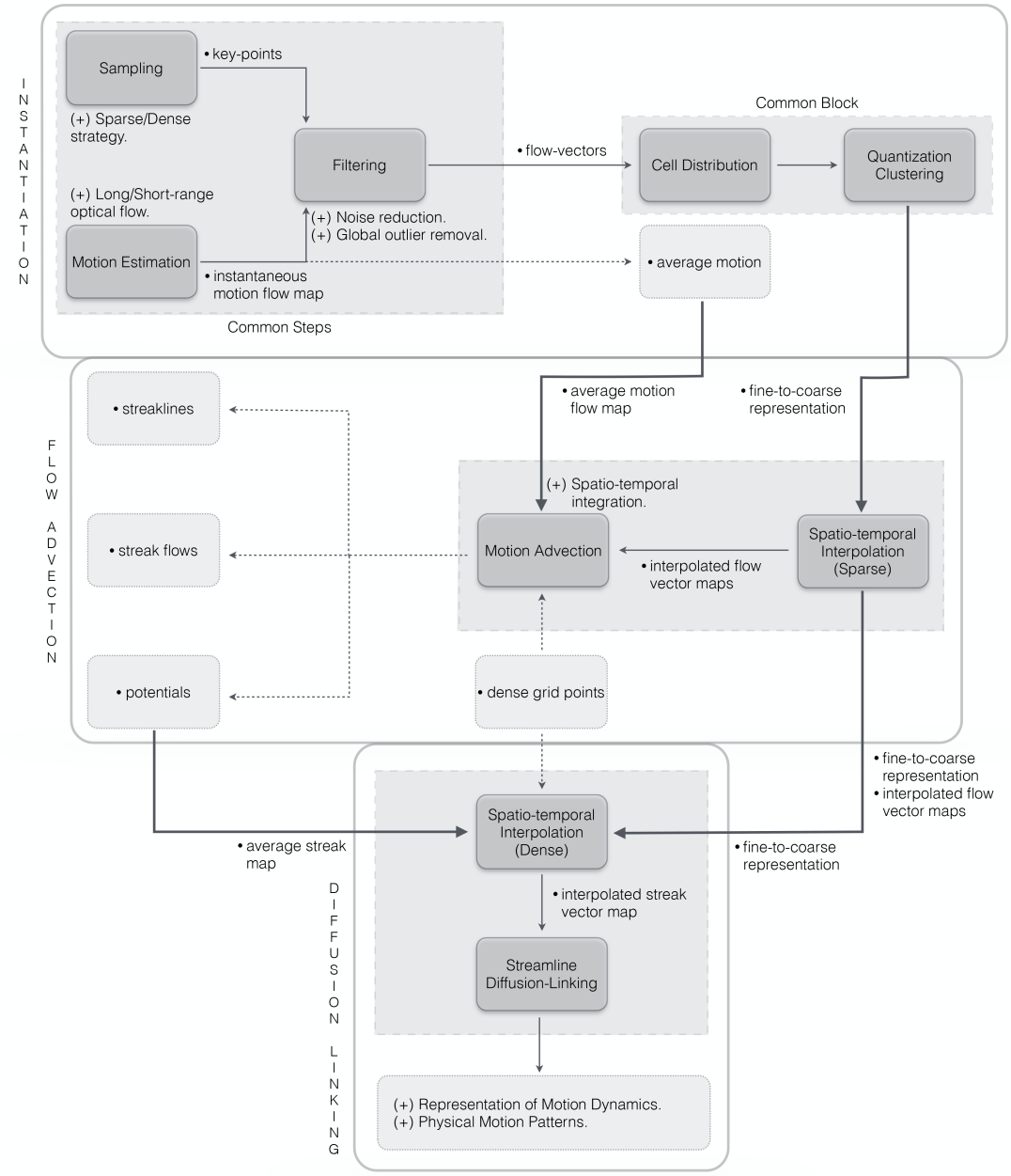


Figure 3.8: Overall outline of ***VILOMA***.

3.4 Validation

3.4.1 Datasets

We present results on several datasets classified according with the following criteria: i) crowd scenario, both structured (C.S.) and unstructured movement (C.U.); ii) multi-tracking scenario, both sparse (MT.S.) and dense groups (MT.D) (see Table 3.1).

Name	Category	Frame Size	Fps	# of Frames
UCF 913-36l [a]	Crowd Structured (C.S.)	480x360	25	467
UMN seq3 [b]	Crowd Unstructured (C.U.)	320x240	30	658
PETS2013 S2 L1 Time12-34 View001	Dense Multi-Tracking (MT.D.)	768x576	–	794
PETS2013 S2 L3 Time14-41 View001	Sparse Multi-Tracking (MT.S.)	768x576	–	240

Table 3.1: Datasets characteristics.

3.4.2 Evaluation Settings

Several experiments were conducted to obtain the parameter's values for a baseline. In this way, the baseline consists in the FAST sampling, the median filtering *kernel size* of $K = (13,13)$, the LDOF's optical flow algorithm [51], the spatial *cell size* of $C = (15,15)$, and the *neighborhood size* of $\mathcal{N} = (3,3)$.

Memory Cell Size	Minibatch Size				
10	2	4	6	8	10

Minibatch Size	Memory Cell Size					
5	3	6	9	12	15	20

Table 3.2: ***VILOMA*** parameters.

Apart of the parameters related to the *instantiation* stage, ***VILOMA*** has two more important parameters: i) *minibatch* size, which sets the length of the streaklines, as well as the flow vectors quantization; ii) *memory cell* size, which sets the number of mini-batches for streamline formation. For further analysis, we vary *minibatch* size fixing *memory cell* size, and vice versa accordingly to Table 3.2. We highlight that the *kernel size* parameter is used for the common kernel operations of ***VILOMA***, such as flow vector computation and filtering, among others.

3.4.3 Outlier Removal

The outlier removal technique assumes a predominant role on ***VILOMA***'s accuracy. Fig. 3.9 provides a qualitative confirmation. A quantitative comparison of the proposed technique with several outlier removal methods is provided in Table 3.3 reported under three metrics, namely true

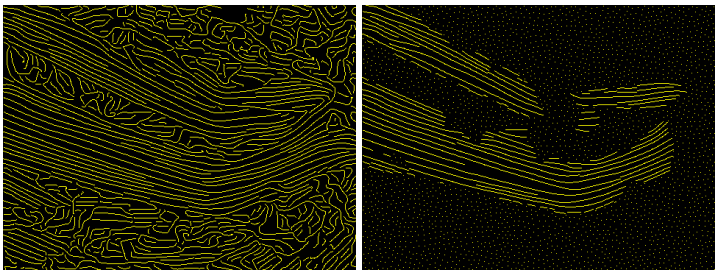
[a] UCF Crowd Segmentation dataset http://crcv.ucf.edu/data/crowd.php
[b] Unusual Crowd Activity dataset http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi

positive (TP) rate, true negative (TN) rate and the mean of both, here called the true balance (TB) rate. Results are computed considering the (C.S.) and (C.U.) scenarios. For the former, nearly 20 frames were masked, since it is a large video sequence with many persons per frame, while for the latter, all the frames were masked. Pedestrians are manually annotated with bounding boxes and the masks are the inner ellipses within each one. The flow vectors that lay outside the masks are considered the background, i.e. the outliers, which are considered as negative samples.

	Std	Thopmson-Tau [350]	Mzscore	Zscore	Adj-Boxplot [137]	ExpSM [165]	Grubbs [101]	Ours
UMN seq3								
TP	97.9 (±1.3)	18.9 (±5.5)	27.4 (±22.6)	**98.7** (±0.9)	58.9 (±10.7)	68.0 (±4.6)	92.2 (±15.9)	79.7 (±6.5)
TN	1.7 (±1.2)	19.2 (±8.1)	16.5 (±6.7)	1.2 (±0.8)	12.7 (±5.5)	21.0 (±4.6)	3.6 (±4.4)	**82.7** (±6.0)
TB	49.8	19.1	21.9	49.5	35.8	44.5	47.9	**81.2**
UCF 913-361								
TP	96.6 (±5.7)	13.8 (±13.6)	29.3 (±33.7)	**97.9** (±3.6)	58.6 (±21.7)	77.3 (±12.1)	76.8 (±31.6)	79.9 (±14.2)
TN	2.3 (±1.8)	24.8 (±21.9)	17.8 (±10.7)	1.7 (±1.3)	21.3 (±21.1)	16.5 (±7.2)	7.8 (±6.2)	**89.7** (±6.4)
TB	48.4	19.3	23.5	49.8	39.9	46.9	42.3	**84.8**

Table 3.3: Sensitivity and Specificity of various outlier removal techniques (%).

We have tested with simple and more complex techniques (for most of them, we use the available source code). For instance, the most *naïve* approach, the *Std* method, considers as outliers the samples that distance from the mean more than 3σ, while more robust approaches follow statistical models that deal with skewed data. All techniques are applied on the log-normalized magnitude distribution of the flow vectors. Results show that our technique keeps a high TP rate and, more importantly, it demonstrates a higher TN rate, proving its effectiveness on removing background's flow vectors. None of the others techniques are able to achieve satisfactory TN rate, which support our evidence to propose a new outlier removal technique for this problem. The combined TB rate clearly shows the overall supremacy of our technique. We conclude that the combination of the Chebyshev's theorem with the skewness metric brings a stabilization factor to the initial rough approximation of the unimodal normal distribution for the formulation of our outlier removal technique.



(a) (b)

Figure 3.10: Effects caused by the outlier removal technique on streamline formation (C.S. scenario): (a) without outlier removal; (b) with outlier removal (single points are seeds that are, correctly, not diffused by lack of meaningful flow field).

Figure 3.9: Comparison results with and without the proposed outlier removal technique. By row: from left to right, with outliers and after outlier removal. By column: from top to bottom, C.S., C.U., MT.D., and MT.S.

This step is also crucial for further motion advection step. Fig. 3.10 confirms the high perturbation that undesired flow vectors introduce on the streamline formation. The great advantage of the proposed outlier removal technique is that it is data-driven, therefore it automatically adjusts to several scene contexts, as well as to large motion variations during the same scenario. We verify that the normal percentage of the removed flow vectors is around $[65\%, 75\%]$, but it could varied around $[20\%, 85\%]$, depending on the scene category and the video characteristics.

3.4.4 Global Motion Trajectories

The main purpose of ***VILOMA*** is to effectively extract long-range global motion trajectories that approximate the ones obtained from multi-tracking approaches, in many different scenarios.

The final trajectories are obtained by the *linking* process. The parameters of the selected farthest point seeding method are fixed to the default values, $d_{sep} = 4.3$ and $d_{rat} = 1.3$, both in sub-pixel accuracy and within the optimal ranges stated in [220]. The former controls the spacing distance of the density field, and the latter expresses the saturation ratio to trigger the seeding of a new streamline. It is necessary to execute a pruning step to remove redundant and short streamlines. Such process removes trajectories with a number of points below the mean value of the whole set of trajectories extracted on all set of mini-batches of the entire sequence. We use in this step the number of points since after the diffusion process the points are very close to each other (nearly one pixel), therefore its value is approximately equal to the trajectory length. Table 3.4 states the significant reduction of the trajectories achieved with the pruning and *linking* processes.

Dataset	Minibatch Size (for memory cell size = 10)														
	2			4			6			8			10		
	BP	AP	AL	BP	AP	AL	BP	AP	AL	BP	AP	AL	BP	AP	AL
C.S.	2058	574	434	918	263	222	573	165	143	426	121	106	405	106	89
C.U.	2508	995	169	1381	540	153	891	346	136	770	283	128	580	196	88
MT.D.	6252	2088	1171	4156	1184	691	3597	1020	566	2695	707	389	2263	616	319
MT.S.	2111	516	247	797	235	127	714	194	128	424	125	87	356	100	60

Dataset	Memory Cell Size (for minibatch size = 5)																	
	3			6			9			12			15			20		
	BP	AP	AL	BP	AP	AL	BP	AP	AL	BP	AP	AL	BP	AP	AL	BP	AP	AL
C.S.	3043	818	598	1372	368	297	879	244	199	618	165	143	497	145	122	347	100	83
C.U.	3514	1398	236	1746	662	175	1180	441	157	877	339	132	716	267	117	535	176	89
MT.D.	3948	1294	775	5705	1656	1001	4167	1178	706	3521	973	552	2949	814	456	2132	570	317
MT.S.	2978	711	351	1260	332	190	741	222	124	704	186	115	347	113	81	367	104	66

Table 3.4: Number of trajectories per framework's parameters on each dataset at different steps (BP: before pruning; AP: after pruning; AL: after linking).

The geometrical constraints for the MRF graph are fixed to $d_{thr} = 45$, $\theta_{dir} = 42^{\circ}$, and $\delta_{dif} = 40^{\circ}$. For the remaining parameters used on the similarities terms, we follow Rubinstein *et al.* [313] with the exception of n_p and n_v, which are the number of points taken for appearance and velocity similarity terms, respectively. They are fixed heuristically considering a percentage of the trajectory with the minimum number of points (n_{min}), $n_p = 0.7 \times n_{min}$ and $n_v = 0.35 \times n_{min}$. The velocity term is lower due to a higher expected variation, therefore less point-to-point measures are considered.

To the best of our knowledge, there is not any evaluation framework that deals with comparison between manual trajectories and automatic long-range motion trajectories of pedestrians. We propose an evaluation methodology that supports clustering of trajectories by similarity, in order to obtain the most representatives and measures the correspondence between extracted and annotated trajectories. We follow a similar approach to Ochs *et al.* [252] and use a one-to-one distance

function between each annotated trajectory and each auto-generated one to obtain a distance matrix, and solve the assignment problem with the Hungarian algorithm [178]. We report the quality of the matching process with the miss detection (FN) and false positive (FP) rates.

The distance matrix passes through a regularization process before applying the Hungarian algorithm. This step favours configurations with small residuals and down-weights large errors that could otherwise dominate the matching process. We evaluate four alternatives to regularize the distance matrix: i) *clustering threshold,* where a K-means is applied to the distances and the *max* of the cluster with the lowest values is taken as the desired threshold to truncate the matrix entries up this value; ii) *quartile threshold*, where the third quartile of the distances distribution is taken as threshold, and a similar truncation process is applied; iii) *median RLS*, where the median of the distances is used as the σ parameter for the robust least square (RLS) approach expressed by $\rho(u,\sigma) = u^2/(\sigma^2+u^2)$, where u is each distance value; iv) *local scaling RLS*, where the same robust least square approach is taken, but instead of using the median value, the mean of a clustering based on local scaling is used. Such value is computed as follows: 1) a K-means is applied to the distances, 2) the *max* of each cluster is used to obtain a distance based on local scaling as stated by Zelnik-Manor and Perona [408], 3) the mean of those distances is considered to be the σ value.

An annotated trajectory is considered correctly matched if its distance to an auto-generated one is below a certain threshold. The threshold is also used to evaluate the relationship between the accumulated error (AE), sum of the distances between the trajectories correctly matched, and the FP rate. Such threshold is considered to be the *max* of the cluster with lowest values after applying the same K-means process detailed in Section 3.3.1.5. Before computing the distance matrix, all trajectories are resampled using a cubic spline interpolation scheme, where the resampling value is global and is equal to the length of the trajectory with the minimum number of points.

Next, we present some results for each dataset that will help to answer the following questions: i) what is the influence of *memory cell* size; ii) what is the influence of *minibatch* size; iii) what is the influence of the regularization matrix distance method on the assignment process; iv) what is the most suitable distance function. We take conclusions based on the relation between the FP rate and the AE, varying the threshold for the incorrectly classified matches. The following results are presented accordingly with the regularization step and the distance function. For instance, the chosen distance metrics between trajectories are Euclidean, Hausdorff, Dynamic Time Warping (DTW) and Longest Common Subsequence (LCS), while the regularization that leads most times to the lowest error is selected. The distance matrix is normalized by the minimum and maximum, therefore the AE for each metric can be compared. The distance functions are calculated considering four feature's trajectory, (x,y,dx,dy), which are the point coordinates and the normalized vector direction between subsequent trajectory segments. Further individual analyses are based on the results reported in Fig. 3.11, Fig. 3.13, Fig. 3.15, Fig. 3.17.

3.4.4.1 Crowd Structured (C.S.) Scenario

For this scenario a manual annotation of pedestrians is performed. Since the motion is structured, i.e. is represented by common motion patterns, we select and annotate several persons that undergo each motion pattern to obtain representatives trajectories. In this case, we obtain 25 trajectories.

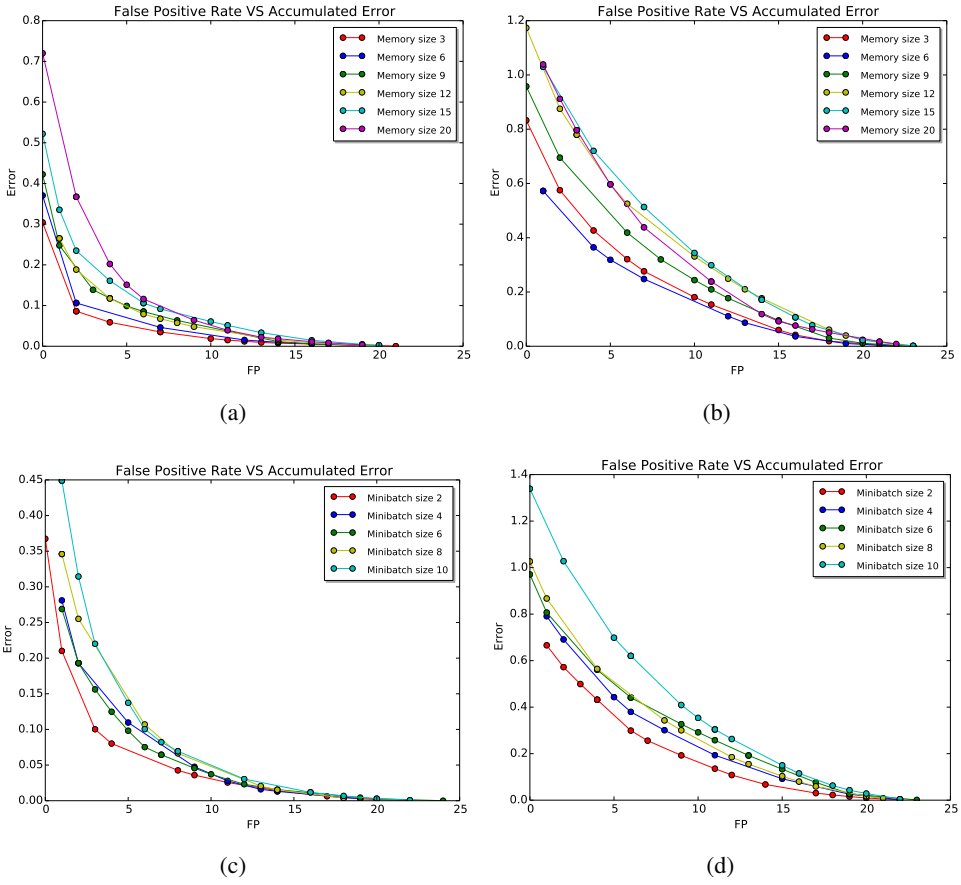
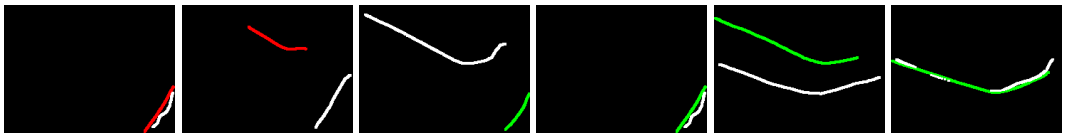


Figure 3.11: FP rate and AE relation on C.S. scenario with RLS local scaling regularization for: (a) *memory* variation and DTW metric; (b) *memory* variation and euclidean metric; (c) *minibatch* variation and DTW metric; (d) *minibatch* variation and euclidean metric.

Both *memory cell* size and *minibatch* size have similar behavior, namely lower values have less error and FP rate associated, therefore for this type of scenario *memory* and *minibatch* size should be kept low. We also notice that *memory* size is less sensitive than *minibatch* size, therefore *memory* size could be higher than the *minibatch* size. The chosen regularization is the *local scaling RLS*, since, in general, it presents lower error. The DTW measure presents a steeper monotonically decreasing behavior which shows a better compromise between the error and the FP rate. It also produces less FPs. Euclidean distance also performs well, while LCS measure presents the worst results, since it has almost a linear behavior (see Fig. 3.11).

Fig. 3.12 presents some matching results, where manual trajectories are represented by white, and auto-generated trajectories are illustrated by red if a FP is considered, or by green if a correct match is assigned. This nomenclature is kept for subsequent figures. Fig. 3.12(a) and Fig. 3.12(d)



(a) (b) (c) (d) (e) (f)

Figure 3.12: Assignment using DTW metric (*minibatch*=2, *memory*=10): (a) false negative; (b) miss-match detected successfully; (c) FP not detected; (d), (e), (f) matches detected.

show the threshold influence when deciding if it is a FP or not, the former is a false negative assignment, and the latter is a correct match. Fig. 3.12(c) presents a wrong match, probably due to the global minimization problem associated with the Hungarian algorithm. However, we state by Fig. 3.12(e) that our framework is capable to generate a possible assignment for the previous miss match. We verify valid assignments and FP detections. It is important to highlight that just some trajectories are annotated for this scenario, which has hundreds of trajectories that could be more similar to the automatic trajectories.

3.4.4.2 Crowd Unstructured (C.U.) Scenario

This is the most challenging scenario for ***VILOMA***. It is a low resolution video that represents a crowded scene where pedestrians move randomly in various directions, causing constant occlusions. We manually annotate all the pedestrians, which lead to a total of 15 trajectories.

In this case, *memory cell* size has a less stable behavior assuming the relation between the error and the FP rate. However, in general lower values present better results. *Minibatch* size has even more variability, and we verify that medium and large values perform better than lower values. The selected regularization is the *median RLS*. In terms of distance function, we reach the same conclusion of the previous scenario, just highlighting the improvement of the Hausdorff metric, which shows a steeper curve on low thresholds, but maintains a large FP rate at higher thresholds (see Fig. 3.13).

In general, Fig. 3.14 permits to take the same analysis as in the previous scenario. We visually verify that ***VILOMA*** extracts trajectories that are very similar to the manual ones. However, the evaluation framework was unable to consider them in the final matching results. This factor leads us to conclude that, despite the larger matching differences, ***VILOMA*** performs well on such demanding scenario.

3.4.4.3 Dense Multi-Tracking (MT.D.) Scenario

This scenario is also very challenging. Normally, multi-tracking approaches are applied to it and flow-based approaches conduct to poor results. Annotation of pedestrians leads to 19 trajectories.

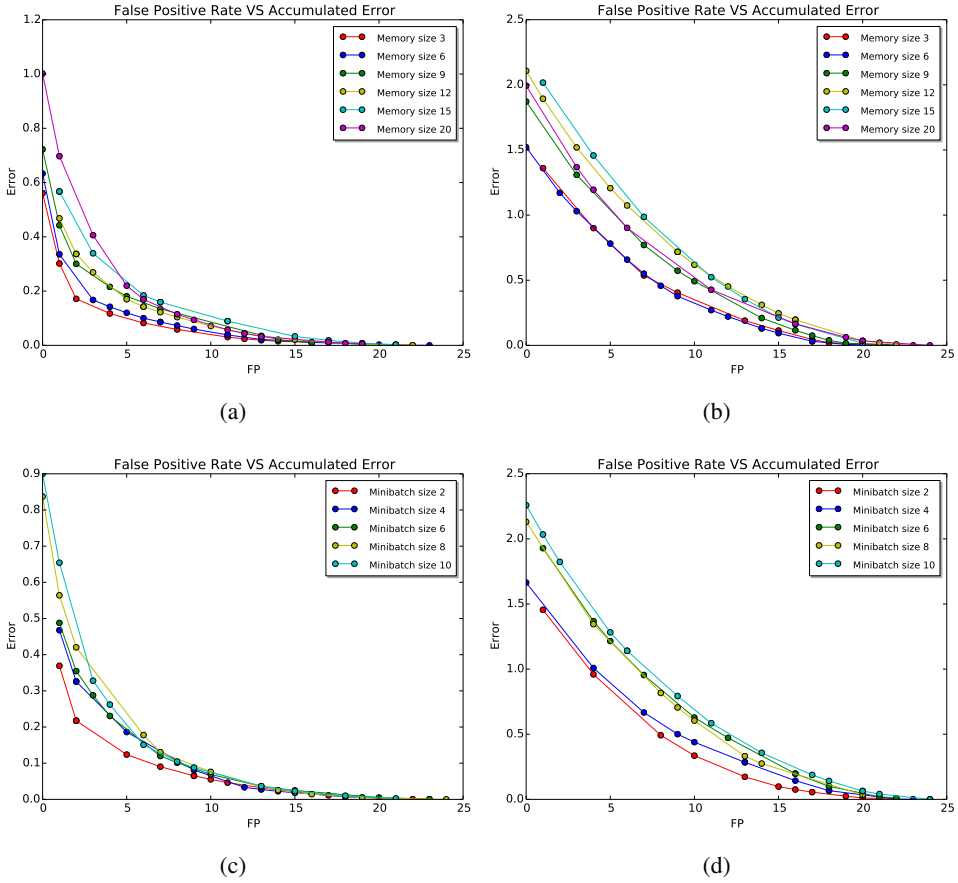


Figure 3.13: FP rate and AE relation on C.U. scenario with RLS median regularization for: (a) *memory* variation and DTW metric; (b) *memory* variation and hausdorff metric; (c) *minibatch* variation and DTW metric; (d) *minibatch* variation and hausdorff metric.

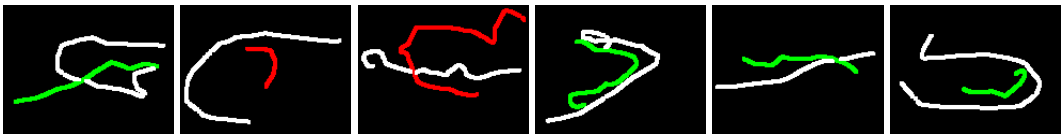


Figure 3.14: Assignment using DTW metric (*minibatch*=8, *memory*=10): (a) FP not detected; (b), (c) miss-match detected successfully; (d), (e), (f) matches detected.

An analysis shows that low values for *memory cell* and *minibatch* size decrease performance, while medium values perform better. The selected regularization is also the *median RLS*. Regarding the distance functions, previous conclusions still hold true. The Hausdorff metric reaches the best performance, which presents a low and steeper error for low FP rates (see Fig. 3.15).

Fig. 3.16 shows the most difficult trajectories to match in this dataset. In fact, Fig. 3.16(a) presents a complex manual trajectory, which is very long and has several direction changes. However, it could be divided on two cross sections, one of them is automatically captured by ***VILOMA***, nevertheless it is not identified as a correct match.

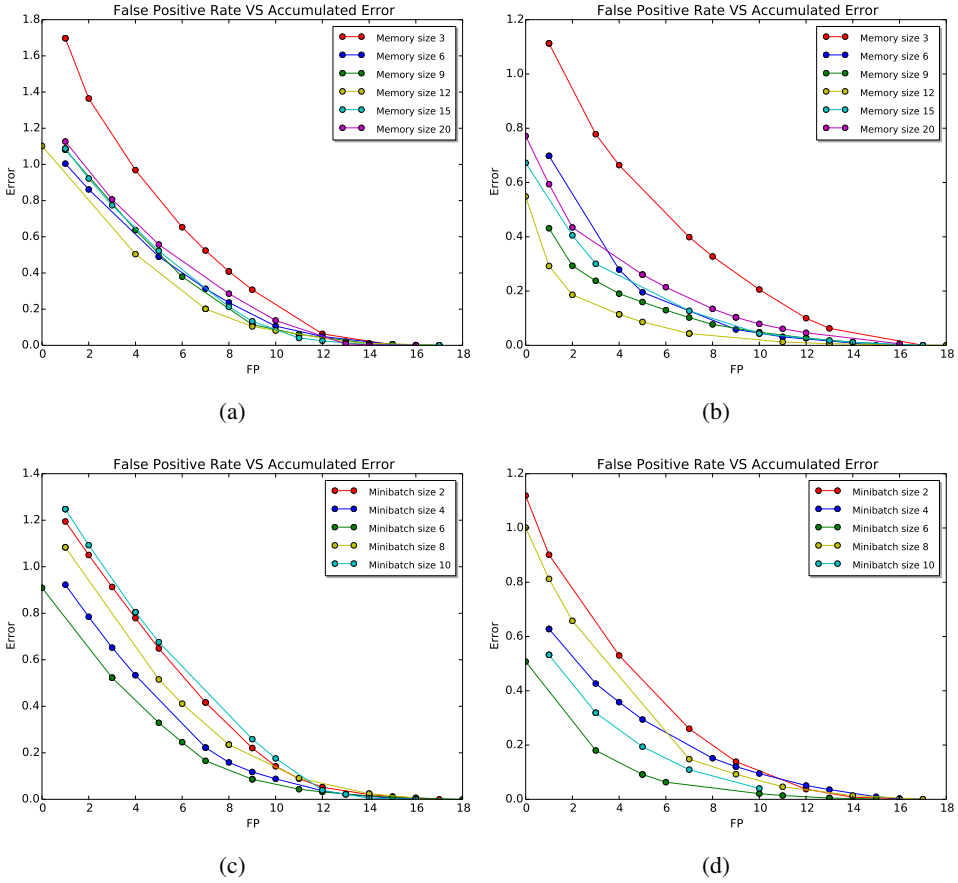


Figure 3.15: FP rate and AE relation on MT.D. scenario with RLS median regularization for: (a) *memory* variation and DTW metric; (b) *memory* variation and hausdorff metric; (c) *minibatch* variation and DTW metric; (d) *minibatch* variation and hausdorff metric.

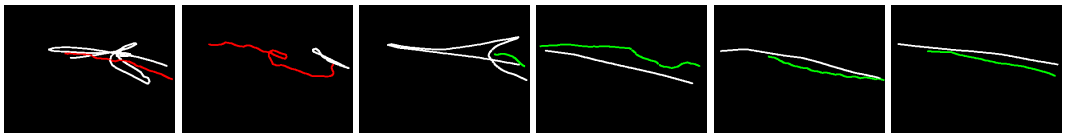


Figure 3.16: Assignment using hausdorff metric (*minibatch*=5, *memory*=12): (a), (b) miss-match detected successfully; (c) FP not detected; (d), (e), (f) matches detected.

3.4.4.4 Sparse Multi-Tracking (MT.S.) Scenario

This is another scenario where multi-tracking approaches are normally applied. This dataset has an additional difficulty related to illumination variations that could lead to erroneous flow calculation. There are 44 manual trajectories annotated from pedestrians.

The performance dependence on *memory cell* and *minibatch* presents the same behavior as in the S.C. scene. The regularization of the distance matrix is based on the *clustering threshold* alternative, and the DTW is, again, the metric which presents the best results (see Fig. 3.17).

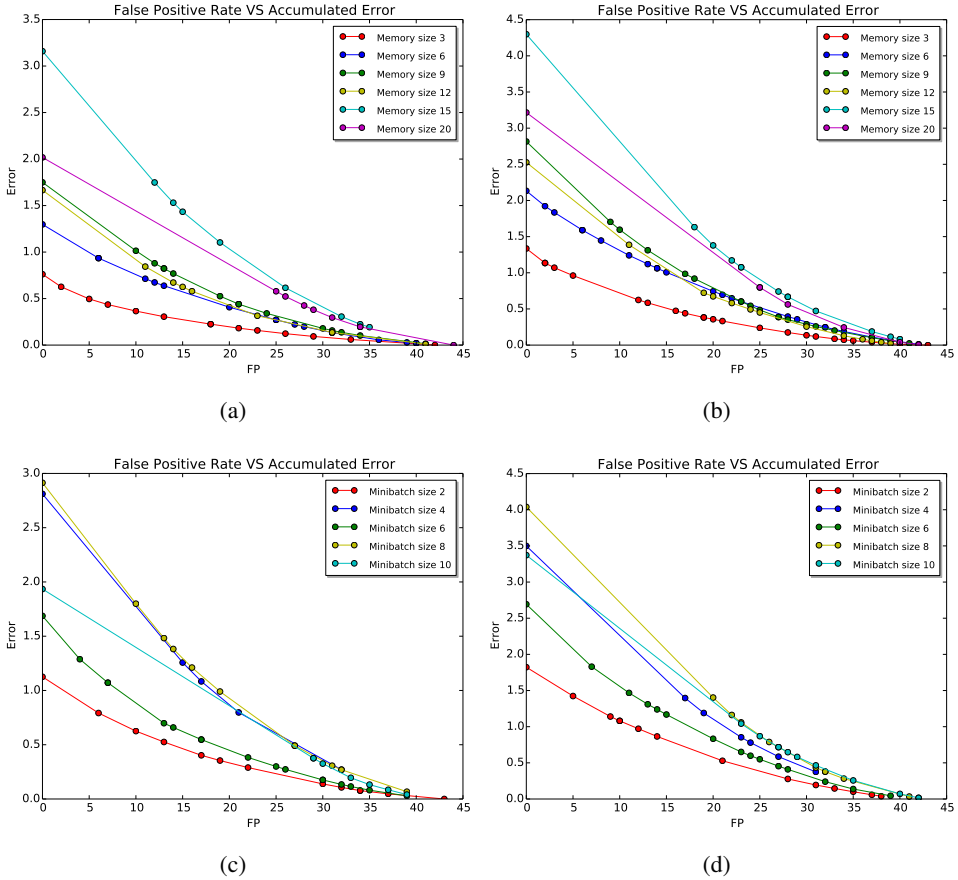


Figure 3.17: FP rate and AE relation on MT.S. scenario with clustering threshold regularization for: (a) *memory* variation and DTW metric; (b) *memory* variation and euclidean metric; (c) *minibatch* variation and DTW metric; (d) *minibatch* variation and euclidean metric.

This dataset presents the best results (see Fig. 3.18), which is an excellent conclusion since it proves the versatility of ***VILOMA***. It correctly captures short and long trajectories, as well as FPs. However, we notice that the matches presented on Fig. 3.18(b) and Fig. 3.18(c) could be inversely assigned to obtain a better matching.

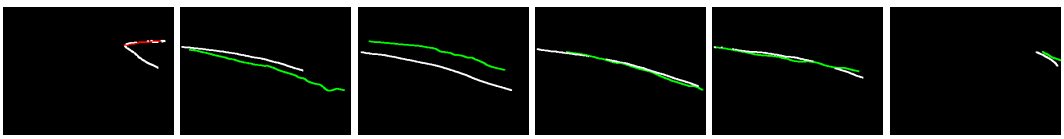


Figure 3.18: Assignment using DTW metric (*minibatch*=5, *memory*=3): (a) miss-match detected successfully; (b), (c), (d), (e), (f) matches detected.

As general evaluation, we conclude that low values should be considered for *memory cell* size, while *minibatch* size can be varied from low to medium values depending on the scene category. Therefore, *minibatch* size presents a larger functional interval, while avoiding error

propagation derived from numeric calculations of the motion advection scheme. The lower values for *memory cell* size permits to capture less linear trajectories, eliminating the possibility to miss the detection of trajectory segments with high curvature, since the *linking* process could favour direction continuity. In terms of regularization, it is obvious to conclude that a non-linear technique improves the results. In terms of the distance function, we verify that metrics that consider point-to-point geometrical information and shape benefit the matching process.

For qualitative validation, Fig. 3.19 presents the spatio-temporal comparison between the extracted trajectories and the manual ones. The density obtained clearly shows the robustness and efficiency of ***VILOMA***. Is important to mention that the assignment process always finds a correspondence between a manual trajectory and an auto-generated one, which corroborates the density evaluation.

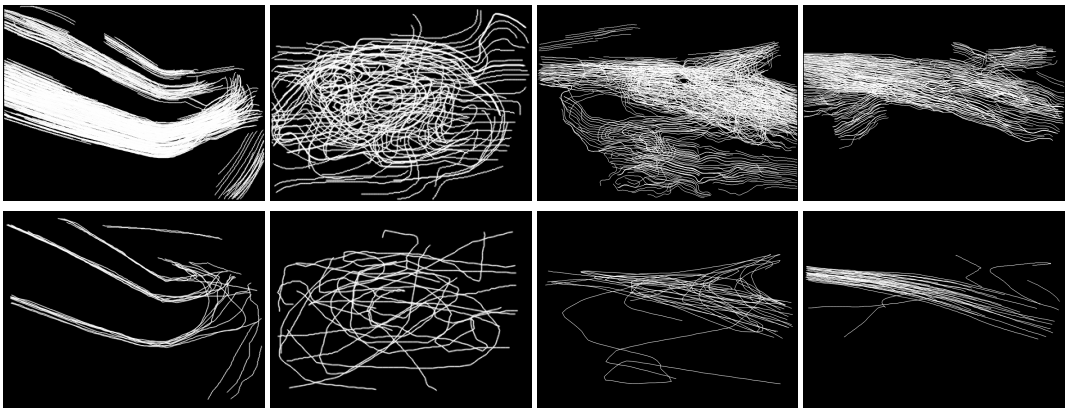


Figure 3.19: Comparison between all automatic extracted trajectories and manually annotated trajectories. Top row: automatic; Bottom row: manual.

3.4.5 Motion Segmentation

This section reports the usefulness of ***VILOMA*** for human activity-related tasks. In this case, we evaluate the motion segmentation performance against two state-of-the-art works [12, 222]. We follow both the qualitative and quantitative analysis approach of Mehran *et al.* [222] to present and compare the dynamic segmentations of different behavior states through a video sequence. For the qualitative analysis, we use the same datasets as Mehran *et al.* [222], namely Argentina and Boston [3]. However, for the quantitative evaluation we do not report the results in the Boston dataset because it is a very time-consuming task to annotate all motion objects (14130 frames). We should highlight that we were unable to reproduce the results in [222], even using the author's source code and the same parameterization [4].

For the experiments, we use the following parameters for each approach: i) pathlines, *integration time*=15; ii) streaklines, *streak length*=40; iii) our, *minibatch*=3, *memory*=5. In order for

[3] http://www.cs.ucf.edu/~ramin/?page_id=99

[4] We acknowledge and thank the efforts of the first author of [222] to try to overcome such discrepancy, although without success.

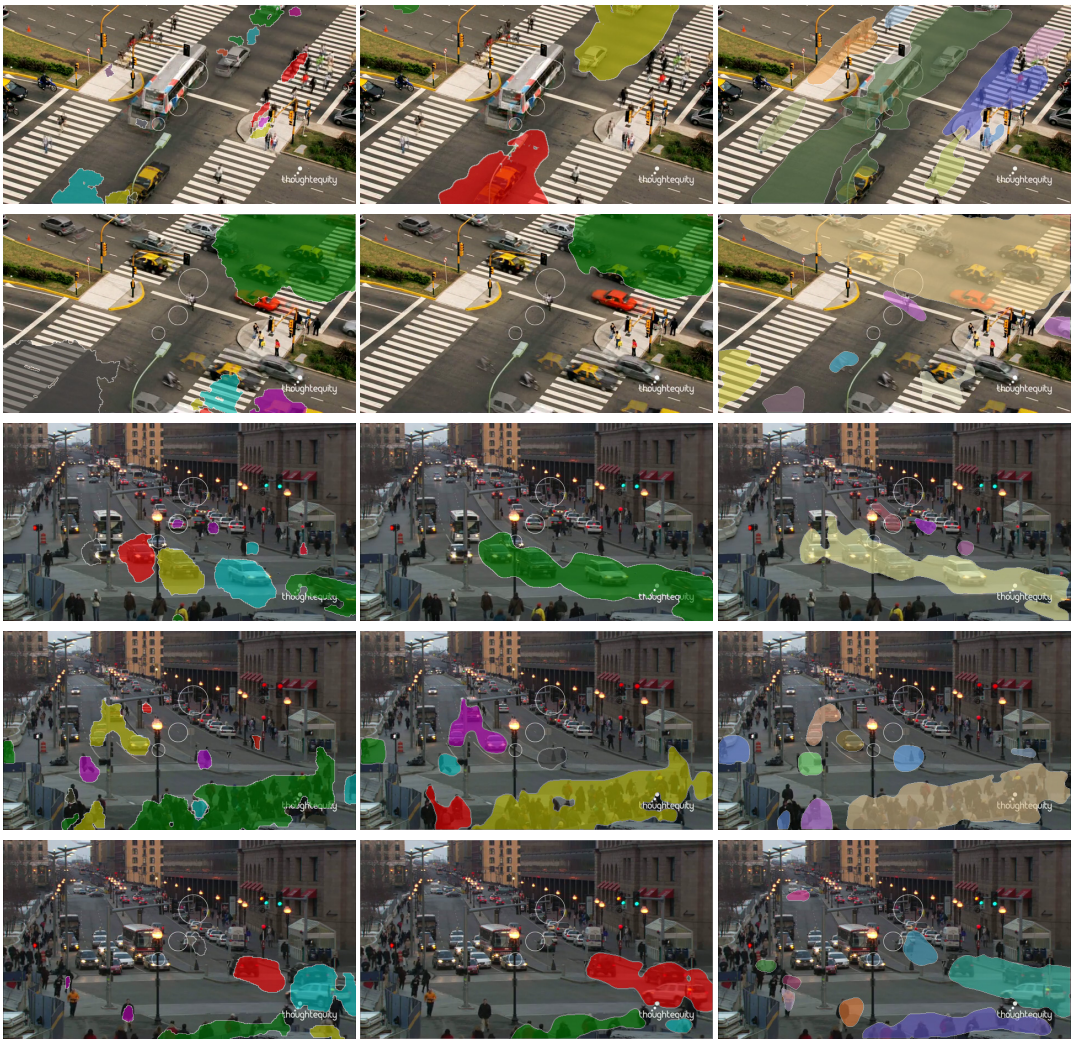


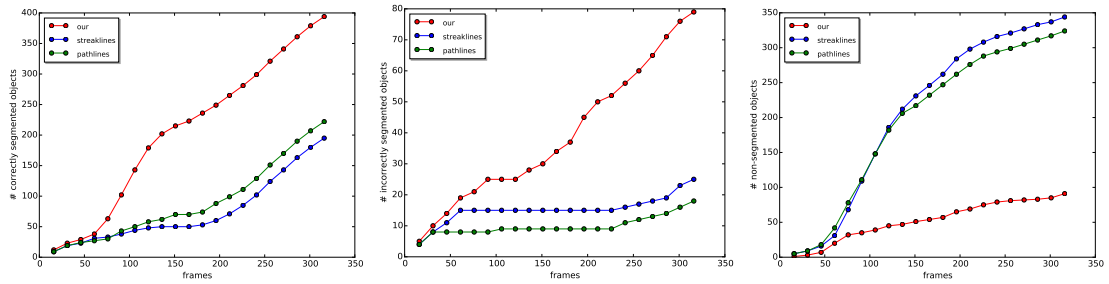
Figure 3.20: Qualitative comparison of segmentation results in Argentina and Boston datasets. By row: frame 115 (1st row), and frame 213 (2nd row) of Argentina; frame 40 (3rd row), frame 433 (4th row), and frame 2042 (5th row) of Boston. By column: from left to right, pathlines, streaklines, and our approach.

the comparison to be as fair as possible we use the same optical flow result per pair of frames for each approach, namely the Classic+NL from Sun *et al.* [344].

Frames 115 and 213 illustrate two behavioral phases in which traffic lights change and north/south flow of pedestrians emerges, and west/east and north/south bound vehicles flow develop. Remaining frames refer to Boston dataset and also consider different behaviors when traffic light changes an east/west flow of pedestrians emerges simultaneously with a north/south bound vehicle flow. Fig. 3.20 demonstrates that streaklines are spatially and temporally pronounced and more accurate on capturing dynamic objects than the pathlines, which show fragmented segments of movement. However, our approach clearly shows longer motions and robustly segments different flows, even on cluttered conditions. For instance, our approach is able to detect flow of pedestrians on each sidewalks and distinguish them from north and south bound (frame 115), and detect and separate

standalone motions (frames 40, 433 and 2042). None of the other approaches are able to do this, even inspecting the results reported by Mehran *et al.* [222]. We clarify that our segmentation approach is computed per pixel, considering the similarity of the cosine difference in a 8-connected neighborhood, i.e. it follows the same process of the streak flow similarity of Mehran *et al.* [222]. In our case, the flow is derived from the long-range trajectories, where each one is resampled and their segments are considered as the flow vectors.

For the quantitative comparison, we also follow the same criterion stated by Mehran *et al.* [222]. Our approach outperforms by a large margin, more than twice, both state-of-the-art approaches in the number of correctly segmented objects, even considering the results reported by Mehran *et al.* [222]. The number of non-segmented objects is less than three times of the reported by the remaining approaches, which corroborates the qualitative analysis. However, it also presents a higher number of incorrectly segmented objects, probably because of an over-segmentation on clutter regions (see Fig. 3.21).



(a) (b) (c)

Figure 3.21: Quantitative comparison of segmentation results in Argentina dataset: (a) correctly segmented objects; (b), incorrectly segmented objects; (c), non-segmented objects.

3.5 Summary

In this chapter, is presented an automatic and novel framework, ***VILOMA***, based on global dense flow and local motion information that extracts meaningful long-range trajectories. It main advantages are its efficiency on different surveillance scenarios, and the correlation between the extracted trajectories and the pedestrians movements. ***VILOMA*** combines a macro and micro analysis of motion, that to the best of our knowledge was not explored before. It presents important contributions on different stages, namely on its formulation, motion extraction and propagation scheme, a novel technique for removal of flow vector outliers, a fine-to-coarse flow representation, and integration of local motion information into a global re-correlation algorithm at the tracklet-level to form long-range trajectories. It is also highlighted the usefulness of ***VILOMA*** in providing robust spatio-temporal motion information to be used for human activity tasks such as segmentation, where it outperforms state-of-the-art algorithms.

An evaluation framework is extended in order to deal with the matching problem between global and individual trajectories. Promising results are demonstrated on different datasets. However, two aspects should be improved: i) the FP rate, since sometimes they are wrongly detected; ii) the assignment process, since in some cases the framework produces most similar trajectories than the ones that were assigned by the matching process, impairing the true performance of the framework. More surveillance streams should be used to optimize the framework's parameters for each scene context. In that way, the framework could get even better results.

The next Chapter 4 presents a new perspective to the analysis of individual and collective activity in surveillance scenarios. It belongs to the same level, *the group*, within our categorization, but it comprises a finer scale analysis. In that context, the *pedestrian* starts to be called by *individual*. The proposed social behavioral analysis framework brings together new deeper abstraction semantics to describe human activity in social context with a trajectory-based representation that encodes relational cues, position-based and attention-based, between individual-individual and individual-scene for further classification of human activity. Both frameworks can coexists, since ***VILOMA*** may provide useful patterns and statistics, such as regions of interest, common paths, etc, that the social behavioral analysis framework needs to build the entities relationships. We believe that detection and classification of human activity could benefit from the framework proposed in this Chapter.

Chapter 4

Individual and Collective Social Behavior Analysis*

> A goal is not always meant to be reached, it often serves simply as something to aim at.
>
> Bruce Lee

The increasing demand for human activity analysis in video surveillance has been triggered by the emergence of new features and concepts to help in identifying activities of interest. However, the monitoring of complex individual and collective human activities that express the sociological context of a scene is a topic not extensively studied in the literature, due to not only its intrinsic difficulty and large variety of topics involved, but also because of the lack of valid semantic concepts that relate human activity to social context.

4.1 Introduction

Spatiotemporal trajectory representations are gaining increased attention in surveillance scenarios to analyze human activity and detect abnormal events. However, the research community has been focusing on solving technical problems associated to multi-tracking techniques and on encoding trajectory-based features to detect individual atomic actions. Some approaches combine scene and object features with trajectory-based descriptors to detect event primitives [299], while others aggregate interactions measures and cues to analyse small groups of pedestrians [107]. None of them explore the integration of scene objects, individuals, and their inter and intra relational features to classify individual and collective social behavior. Modeling human activity within a sociologically principled way has an undeniable value for both low-level problems such as pedestrian tracking, and high-level applications such as anomaly detection in security and human behavior prediction for marketing purposes.

*Some portions of this Chapter appeared in [280–283]

This Chapter presents our analysis of motion and relational-contextual features to model the individual and group behavior as social entities in surveillance settings, namely a shopping-mall, where the relationships between the individuals within the same group and with the objects of interest of the scene assume a relevant role. Pure motion-based frameworks, as the one presented in the previous Chapter 3, are not appropriate to model, by themselves, the individual and group behavior. However, they are suitable to gather local and global useful information, and transform them into rich representations to improve the knowledge about the relationships between the individuals and the scene. To accomplish a fully automatic system that analyze social behavior of people in surveillance settings, both type of analysis must coexists: i) *coarse-contextual*, the one that, at a coarse-level, extracts contextual information from scene through the analysis of motion; ii) *fine-relational*, the one that, at a fine-level, recognizes the entities (individuals, groups, objects) of interest in the scene, their attributes, and creates a model of relationships among them. Both analysis embrace our first level of human activity, namely *the group*.

We propose new concepts of individual and group behavior that consider environment context to address the topic of social semantic meaning, and a complete framework, so-called **VIdeo-based SOcial Behavior Identification** (***VISOBI***), that encompasses several factors: a relational descriptor that emphasizes position and attention-based characteristics, automatic features extraction processes, including an improved group discovery algorithm, two feature relevance analysis techniques, and a novel two-fold classification approach based on mini-batches. Special attention is given to the evaluation process that outlines the performance impact of automatic features extraction processes into the classification framework, the analysis of the sociological meaning of the individual features, and the inspection of the features importance while enhancing the class inter-separability to improve the classification rate.

4.2 Overview

4.2.1 Feature-Relevance Analysis

Video-based human activity recognition is a very demanding task that extracts a very large set of features from diverse sources, such as motion, appearance, texture, among others. Such information normally presents high dimensionality, therefore it can benefit from a feature selection process to select the most prominent features, while keeping or increasing the overall discriminative power of the recognition system.

Most of the state-of-the-art methodologies rely on variability based relevance analysis methods like PCA to select and embed the most relevant features from the original representation space. For instance, Shao *et al.* [331] employ a Bag of Words (BoW) model based on Spatio-Temporal Interest Points (STIP) features, and the set of descriptors are later reduced in dimensionality by a round of Principal Component Analysis (PCA) capturing 95% of variance. Vrigkas *et al.* [361] use trajectories of tracked optical flow motion features along with PCA to reduce the dimensionality and obtain the highest performance, while speeding up the behavior of the algorithm. Siddiqi *et*

al. [333] use StepWise Linear Discriminant Analysis (SWLDA) to select localized features and discriminate their class based on regression tests. Molina-Giraldo *et al.* [231] incorporate multiple information sources into multiple kernel representations to enhance the data separability into the scene. A new discriminative supervised method combining the classical PCA with the correlation criterion for performing activity recognition in a smart home is presented by Fergani *et al.* [97]. Nonetheless, in most of the cases, enhancing the variability among samples does not translate into class inter-separability. A different approach is proposed by Ribeiro *et al.* [304], where a bayesian classifier along with a brute search of a subset of features and the relief algorithm are applied. However, the experiments are carried out for the classification of activities which are not very demanding, and the scenarios are highly controlled.

4.2.2 Trajectory-based Descriptors

Trajectory dynamics provide intrinsic features that can be used to build useful representations to analyze several application-driven interests such as scene topology, event detection, social interpretation, and activity classification.

A common practice is to use trajectory information to model a scene by a topographical map composed of nodes, which are the areas of interest, and edges, which represent the connectivity between those areas and encode the activity of a human. Makris and Ellis [208] classify the areas of interest as entry/exit zones, junctions, intersections, and stop areas, which are defined by trajectories characteristics. Pusiol *et al.* [299] define the trajectory slow points as individual topologies which are combined to form the general topology. After that, they segment trajectories by topology affinity building a descriptor composed by primitive events. In general, the statistical and geometric structures inferred from the scene model could be used in a feedback loop, for instance they could filter out false detections or could enrich tracking approaches that incorporate scene context.

Trajectory information could help to detect typical and unusual events. Owens and Hunter [262] apply a self-organising feature map neural networks, with trajectories encoded as point-based flow vectors, to learn normal trajectories and detect new event-related trajectories. However, they cannot distinguish between new normal paths and abnormal behaviors. Khalid and Naftel [163] solve this problem and state an accuracy improvement for trajectory classification and event detection when Fourier coefficient space is used instead of trajectory space. However, tests are just carried on synthetic and manually annotated data, and as complex trajectories are not suitable to a global Fourier approximation, probably this approach would not perform as good on real scenarios.

The sequence of trajectories' characteristics are normally used to extract motion patterns that segment the scene into semantic regions. Wang *et al.* [373] introduce a clustering algorithm that accounts with similarity measures and comparison confidence between trajectories to obtain clusters of different activities. This type of approaches largely depends on the distance measures, which also vary depending on the activities being detected. To overcome those errors, some approaches formulate the problem in a probabilistic way. Wang *et al.* [372] propose a nonparametric

bayesian framework. The number of clusters for both the observations of an object on a trajectory and the trajectories are simultaneously learned from a dual Hierarchical Dirichlet Process (HDP). It also permits to reduce the space complexity. However, since trajectories are modeled as words to be quantized into a codebook, such representation lacks of temporal information.

4.2.3 Group Dynamics

An extension of activity analysis embeds notions of social psychology, normally applied to discover and characterize groups of people. Relational connections among people, focus of attention of each person, geometric scene constraints, and proxemics-based distances are ones of the cues used on this type of analysis. Normally, such information is used by microscopic approaches, whose can be divided into social force model-based [126], virtual agents [170], and cellular automata [13]. In particular, Chang *et al.* [63] adopt a probabilistic grouping strategy which accounts with a pairwise spatiotemporal measure between people. A connectivity graph is built for further segmentation of groups and derivation of individual probabilistic models. Each model considers motion type, related to atomic action, direction distribution and distance change, related to interactions. However, no object-scene relation is considered, and they do not intend to use relational context to describe individual profiles. In fact, common microscopic approaches try to simulate pedestrian physical behavior and infer characteristics about group formation, dispersion, and evolution, but they do not capture individual semantics.

The problem of analyzing groups with temporal reasoning implies the solution for automatic group discovery, its formation and dispersion. Recent approaches on group discovery mainly focus on F-formations [72, 138, 328, 409], a particular group type composed by congregating people with the intent of chatting and exchanging information among them. These techniques assume a static environment with little movement of the people involved, mainly focusing on the spatial arrangement of the group (e.g., vis-a-vis, L, side-by-side, circular). The most important feature is the o-space, a convex empty space surrounded by the people involved in the social interaction, in which every participant looks inward. However, some recent works focus on more complex situations. For instance, Bazzani *et al.* [30] consider two dependent subspaces where individuals and groups share the knowledge of the joint individual-group distribution. This work outlines the benefits of defining the dynamics of the group given individual information and vice-versa. Aimed group processes are merged, split and queued. Likewise, Zaidenberg *et al.* [407] address a more flexible group discovery based on features like distance, speed and orientation of the participants, aiming at providing a comprehensive solution for increased group dynamics from typical video surveillance environments, i.e., group creation, update, merge, split, and termination. However, their evaluation is not conclusive, since it was conducted over just three videos with a reduced number of frames.

4.3 Semantics Concepts

The annotation of human nonverbal behavior should reveal meaningful representations semantically associated with ontological concepts for human activity. The diversity of theories that intent to explain and measure the linkage between the psychophysiological states and the human behavior has been triggering different representation approaches on the literature that account with temporal process of actions, spatiotemporal relationships between entities, poses, gestures, among others [172].

A general view about different semantic levels used on human activity analysis is presented by Aggarwal and Ryoo [3]. They define a hierarchical approach where semantic levels are related to an increasing complexity of human activity categorization: i) gestures, elementary movements of body parts such as *raising an arm*; ii) actions, atomic activity composed by temporal sequence of gestures such as *jumping*; iii) interactions, a sequence of single activities between two or more persons such as *a person hugging another*; iv) group activities, single or complex activities performed by a conceptual group such as *a group having a dinner*. Such levels follow an analogy to grammar-based semantics that can be used to map annotation labels to relational inferring models, for instance an action is associated to a verb, a gesture to a phrase where the entity is the body part, etc. Some works have already defined semantic concepts: division between the part of actions as objects and the poselets closely related to those actions [402], Allen's temporal predicates [15] applied to features and entities to model activities with complex structure [317, 403], definition of topological and directional relations between persons to build context-free grammars [151] and collective context descriptors [66].

Our aim is to add and explore semantics for Individual Profiles (I.P.) and Group Behaviors (G.B.), a topic which is under-explored in the literature. In terms of I.P.s, we follow a grammar-based analogy and present an abstraction layer that can be associated to adjectives, since we are qualifying person's behavior characteristics. In terms of G.B.s, we adopt the definition of group dynamics presented by Cartwright *et al.* [57] that explains the interdependence degree among individuals and their influence over the group behavior they belong to.

All the individual and collective behaviors are characterized considering the environment as social context. The following I.P. concepts are defined:

- *exploring (Exp.)*, when no specific interest is revealed, but movement and gaze are coherent with the scene structure and context;
- *interested (Int.)*, when an interest by an object in the scene is explicitly revealed;
- *distracted (Dist.)*, when no specific interest is revealed which translates into unstructured movement and variability of gaze;
- *disoriented (Dis.)*, when confusion concerning interests is revealed, expressed as high variability of movement and gaze along with an unstructured movement.

In terms of G.B. concepts, the following are identified:

- *equally interested (E.I.)*, when a group presents a coherent behavior, i.e. one of the following conditions are satisfied: i) individuals show interest for the same object, therefore all I.P.s should be *interested*; ii) individuals explore the environment in a similar perspective and in a close position, therefore all I.P.s should be *exploring*, their gaze should be similar, and they should be close to each other.

- *balanced interests (B.I.)*, when individuals within a group do not reveal the same level of interest but maintain the same behavior, i.e. the following condition is verified: i) individuals explore the environment in a similar perspective but not so close to each other, therefore all I.P.s should be *exploring*, their gaze should be relatively similar, and they can be slightly separated from each other.

- *unbalanced interests (U.I.)*, when a group reveals different types of behavior in the scene at the same time, i.e. the following condition is satisfied: i) individuals show different individual profiles and the distance among them, and their gaze can vary.

- *chatting (CHAT.)*, when a group can be considered a free-standing conversational group (FCG), i.e. the following condition is satisfied: i) individuals should be fixed in a position talking with each other (moving individuals while chatting are not considered). By default, all the I.P.s are considered as *distracted*.

4.4 Dataset and Annotation

We selected the Israel Institute of Technology (IIT) dataset and were granted access by the authors [1]. The dataset is composed of several real-life surveillance scenarios such as shopping, the subway, and the street, which provide large duration, intense activity, and high diversity of semantics in terms of individual and collective activity. We chose the shopping-mall since its context provides well-defined social behaviors. This scenario comprises three videos (resolution 512×384 @25 fps) with duration: 83155 frames (55‘26“), 59969 frames (39‘58“), and 90525 frames (60‘21“), but, at the present time, and due to the intensive manual labour involved, only one video has been annotated (the first one). The dataset, including our annotation, is available upon request.

We were advised by staff of the lab of social-psychology of the University of Porto[1] during the annotation process. They helped us to analyze and identify the I.P.s and the G.B.s. Such identification was done with the knowledge of social influence, perception and interaction concepts based on social context (namely culture and physical space). We validated the annotation process considering the sociological objective measure proposed by McPhail and Wohlstein [219], but a complete validation in the field of social-psychology would require an intense and continuous observation process of the space. This effort represents a completely new methodology for social annotation of datasets in the field of computer vision.

[1] Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto - http://sigarra.up.pt/fpceup

# Frames Annotated	Annotation duration	Elapsed Time (s) (I.P.)		Elapsed Time (s) (G.B.)		I.P.s distribution		G.B.s distribution		Average Individuals per frame	Average Inidividuals per group
80894 (97.3%)	02:22:49 (hh:mm:ss)	203.5 35.3 12.8 4.2	Dist. Exp. Int. Dis.	30.7 23 100.3 83.7	E.I. B.I. U.I. CHAT.	45 776 41 7 869	Dist. Exp. Int. Dis. Total	193 27 28 7 255	E.I. B.I. U.I. CHAT. Total	3.5	1.8 (*max*: 9)

Table 4.1: Dataset statistics.

Table 4.1 summarizes some relevant statistics about the annotation, which is subdivided into two levels: i) *low-level features*, related to human detection and tracking, trajectories are acquired from a bounding box enclosing an annotated person on each frame. Re-identification is not considered. When a person is strongly occluded (approximately more than half of the body), his/her bounding box is not marked. Also, a full-oriented gaze-direction $[0°, 360°]$ is annotated over the person's head. Objects of interest in the scene are marked, namely candy box, toy cars, and electric stairs (see Fig. 4.1(b)); ii) *high-level semantics*, related to I.P.s and G.B.s labels, where a trajectory and a group of trajectories reveal different profiles and behaviors respectively. Group formation and dispersion are also marked.

Since we are dealing with position and attention-based features, the trajectories should be projected onto the ground plane to correctly estimate distances and angles of interest. Such a transformation involves camera calibration and geometry reconstruction steps which are briefly described in next section.

4.5 Camera Calibration and Ground-Plane Projection

This stage assumes a relevant role to correctly achieve directional and geometric information. We first obtain the camera parameters through calibration, and then estimate the vanishing lines over the rectified image. Both steps are combined sequentially to obtain the ground-plane projection.

We take advantage of a chessboard that appears on some frames, as shown in Fig. 4.1(a), to get the image and object points to proceed with the camera calibration. We mark out the initial and final frames from which the chessboard is visible, and run a simple template and cross-correlation matching, with rotation and scale invariance, to detect the chessboard on those frames. We only use as image points the outer corners points, since the inner rectangles are not equal and are not aligned. In order to get relevant samples and better calibration results, we only take the frames where the points suffer a significant translation or rotation. To obtain an approximation of the real object points, we relate the physical proportions of the chessboard with the dimensions of a normal person. From this calibration pattern, we extract the camera's intrinsic and extrinsic parameters, as well as the undistorted and rectification matrices to apply on video frames.

To compute the vanishing points, ($\mathbf{v}_x$, $\mathbf{v}_y$, $\mathbf{v}_z$), we adopt the M-estimator SAmpling and Consensus (MSAC) algorithm [248]. Its input is a set of line segments automatically detected using

(a) (b)

Figure 4.1: (a) Detected chessboard points for camera calibration; (b) Horizontal vanishing line (blue), ground plane's projection area (green), ground points (red) to calculate scale factors and reprojection errors, and objects of interest (purple).

a combination of the Probabilistic Hough Transform (PPHT) [214] with the Line Segment detection using Weighted Mean-Shift (LSWMS) [247]. For the former one, we use several incremental threshold levels, and set up a minimum line length and a maximum allowed gap between points on the same segment. These parameters reduce the erroneous lines normally detected in noisy regions. For the latter, we turn off the weighted Mean-Shift step, since it does not improve accuracy and increases computational time, and rank out the line segments by orientation error and kept the highest K. We define a maximum number of segments, S, to detect. Since the results of LSWMS are normally better than the PPHT, we use more segments generated from LSWMS, setting $K = \omega S$ (empirically $\omega = 0.8$). For the remainder, we only consider the results from PPHT that correspond to spatial regions where no valuable information from LSWMS is obtained. This combination allow us to obtain reliable and complementary line segments which are spatially distributed on regions with edge information.

Following the work of Criminisi [71], we identify the vertical vanishing point, $\mathbf{v}_z$, and compute the vanishing lines for the planes in the scene. These geometric cues provide the direction of lines and the orientation of planes, and an image-to-world mapping for each image point that could be calculated, the so-called homography matrix H. Indeed, considering the ground plane, $z = 0$, a plane-to-plane mapping can be defined, $X^T = H^{-1}x$, where x is the image point, X is the ground point and $H = [a\mathbf{v}_x \quad b\mathbf{v}_y \quad \mathbf{l}]$, where a and b are scale factors, and $\mathbf{l}$ is the normalized horizontal vanishing line.

The cross-ratio invariance concept [71] is used to compute the ground plane scale factors as well as the scale height factor. For both, physical measures should be known. In our case, we estimate them by physical relationships. For the height factor, we consider a linear trajectory that passes through the farthest ground plane segment to the closest one, to include perspective distortion, and considered a mean human height ($\approx 1.75m$). For the ground plane scale factors, we use the rectangle points identified in Fig. 4.1(b) and approximated their measures. For each point,

their two components are calculated on the ground plane coordinate system defined by $\mathbf{v}_x o \mathbf{v}_y$. Since these points are aligned, we explore their geometric pattern to compute re-projection errors. To obtain the gaze angle in the ground plane, both the gaze direction vector's points are projected and the angle is measure considering the defined ground plane coordinate system.

To obtain the ground plane's projection area, the user should indicate three points: the origin, a collinear point, where both define a segment, and a point located on the parallel and opposite segment. Using these points and the geometric information previously discussed, the projection area on the ground plane is automatically determined (see Fig. 4.1(b)).

4.6 Social Behavioral Analysis Framework

Aiming to develop a methodology for the automatic detection and identification of social behavior dynamics, we propose a complete framework that incorporates low-level processing modules such as pedestrian detection and tracking, and gaze estimation, conducts mid-level processing related to group formation and dispersion, extracts relational features and elaborates a descriptor representation. The semantics label detection and classification of individual and collective activity, I.P.s and G.B.s respectively, can be pursued in a two-fold manner by: i) the encoding of the relational descriptor into a Bag of Features (BoF) representation, whose workflow is illustrated in blue in Fig. 4.2; ii) a feature relevance analysis process that finds the set of relevant features from the relational descriptor, and use them as a feature embedding tool to find new representation spaces in order to improve the classification performance and avoid the calculation of unnecessary features, illustrated by green in Fig. 4.2. It is important to emphasize that the latter procedure does not undergo the BoF, as stated in the general view of ***VISOBI***, exposed in Fig. 4.2.

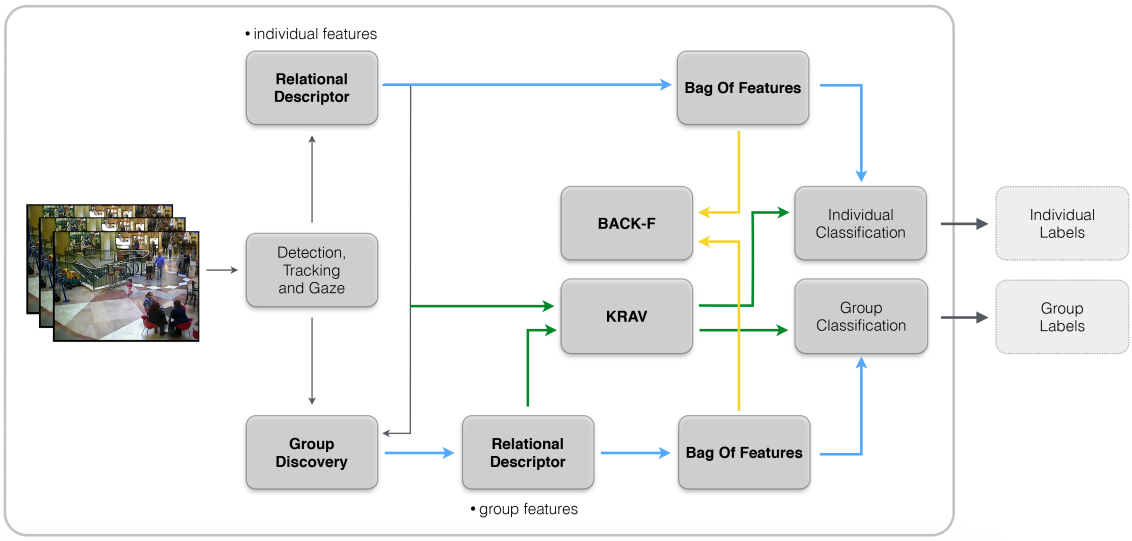


Figure 4.2: ***VISOBI*** framework for I.P. and G.B. identification.

4.6.1 General View

VISOBI is processed frame-by-frame and extracts temporally different features through the video. The scene information, namely the objects of interest, is known *a priori* from the manual annotation. Given a new input frame, pedestrian detection and tracking, as well as gaze estimation, are conducted to automatically obtain trajectories and view orientations, respectively. Both informations are used to extract relational features, position and attention-based, between individuals and objects of interest in the scene. The individual trajectories are also employed to automatically discover group formations and dispersions, while maintaining the structure of the detected groups along a temporal window. Relational features from gaze and direction of movement of the individuals that belong to the same group are extracted. Both individual and collective relational features are then encoded by the relational descriptor. The framework permits to feed the G.B. classification process with the I.P. labels, previously predicted.

Afterwards, the relevance feature analysis process is conducted to obtain the most relevant features, and inspect the importance of the subset of the relevant features, and can also be used to improve the final classification results under challenging conditions such as high imbalance data. Two different techniques are investigated: i) *BACKward-Feature selection*, so-called *BACK-F*, which measures each individual feature importance in the BoW representation, allowing to take conclusions about their sociological meaning; ii) *Kernel-based Relevance Analysis for Video data*, denominated *KRAV*, which embeds the set of relevant features to improve final classification. It also permits to elaborate conclusions about social feature importance.

To the best of our knowledge, this is the first unified methodology that solves simultaneously the identification of social dynamics for individuals and collective behaviors in such a demanding scenario. The proposed framework, so-called ***VISOBI***, encompasses a complete pipeline that undergoes different layers of processing in a bottom-up perspective, with an embedded feature selection classification approach. The knowledge of the scene and the accurate discovery of group structures permits to explore new relational features that express human activity based on interactions among humans and the space. Such features are encoded into a multi-resolution histogram descriptor that represent them at different granularity levels. It allows large subsets of features' combinations to be explored in terms of their relevance and mutual importance, in order to describe complex semantic concepts related to individual and collective human activity.

4.6.2 Pedestrian Detection and Tracking

For automatic tracking of pedestrians a feasibility study is conducted, aiming to identify requirements and potential limitations. Several promising state-of-the-art algorithms are considered, such as Multiple Instance Learning (MIL) [24], Boosting [112], MedianFlow [159] and Track Learn Detect (TLD) [158]. Among these, only two are actually evaluated (Boosting and MedianFlow), given the technical issues of the available implementation of MIL (memory leaking) and the unsuitability of TLD for video surveillance scenarios. Boosting has the advantage of on-line training and its trade-off between performance accuracy and computational time is controlled by the

features used on the appearance model. For its turn, MedianFlow bases its contribution on the penalization of inconsistent trajectories taken from forward-backward error propagation.

After empirical experiments, Boosting algorithm proved to be the best tracker. However, due to the scene complexity, individual tracker's performance might not be accurate enough. In this way, several improvements are considered, named here *Boosting-Improved*. A motion model based on a kalman filter is embedded, which helps to predict how people will move. The automatic pedestrian detection proposed by Molina-Giraldo *et al.* [232] is performed, which employs background subtraction techniques to restrict the searching only over the foreground regions. Such detected regions are used as hypothesis for the tracking algorithm in the update step. Finally, the multi-assignment problem was weighted with i) an *appearance* term, based on HOG and haar-like features; ii) a *geometric* term, based on the bounding box area and aspect-ratio; iii) *kinematics* term, based on distance and velocity measures; iv) *confidence* term, based on the score evidence given by the pedestrian detection technique; and solved using the Hungarian method [178].

These tracking algorithms are integrated into ***VISOBI*** and suitable metrics are adopted for performance comparison, Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA), from [34].

4.6.3 Gaze Estimation

For automatic gaze estimation the method presented by Chamveha *et al.* [61] is adopted. It relies on unsupervised learning of head orientations preceded by several preprocessing tasks: i) tracking with the head detector from [297]); ii) walking direction estimation (polyline simplification using the Douglas-Peucker algorithm, and line fitting); iii) outlier segment rejection rules; iv) selection of representative images through Mahalanobis distance; v) oversampling for handling imbalanced data. For the tracker, we select the Boosting-Improved tracking algorithm, and for the head detection we integrate the fastHOG library [297] into ***VISOBI***. The fine tuned parameters of the head detection technique are summarized in Table 4.2, where ε is the threshold for the Douglas-Peucker algorithm, and τ_n, τ_l, τ_{var} are the thresholds for rules no. 2, 3, 4, respectively, for outlier segment rejection [1].

ε	τ_n	τ_l	τ_{var}
20	0.5	10	2750

Table 4.2: Recalibrated parameters of the method of Chamveha *et al.* [61] for our scenario.

Fig. 4.3 presents the eight head orientations (classes) considered for gaze estimation and an example for each class. More representative images, which result from the preprocessing steps and further used as training data, are illustrated in Fig. 4.4. Note that the head images are converted to grey scale, normalized and resized to 20×22 pixels.

[1] We thank to the first author of [61], Isarun Chamveha from the Institute of Industrial Science – The University of Tokyo, for helping us to recalibrate the technique for our dataset.

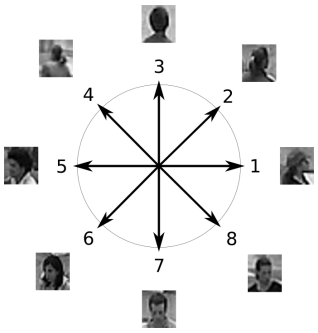


Figure 4.3: Head poses divided into discrete classes for gaze estimation.

Figure 4.4: Representative images generated by the method of Chamveha *et al.* [61]: a row per class in the exact order of the classes from Fig. 4.3.

4.6.4 Group Discovery

Our group discovery approach is based on the work of Zaidenberg *et al.* [407], whose definition of group states that *"a group is a set of mobiles having similar locations, directions and speed for a noticeable duration of time"*. Several reasons take us to prefer this approach among others in the literature, namely the preliminary processing steps of this algorithm are similar to the ones presented in our proposed methodology [280,281], the conceptual definition of group that it stands for is tied with the definition of group dynamics in [57], which we have used to define the G.B. concepts, and the coherence measure encompasses position-based features that are close to the ones that we have employed in our descriptor. However, it also presents some drawbacks regarding

Term	Description
temporal window	time delay to robustly detect and track groups of people (e.g., 10 consecutive frames).
mobile	bounding box of a tracked person at a given frame.
best mobile (of a group)	the mobile from the group that is the closest to its center of gravity.
father / son (of a mobile)	an older / more recent (by frame number) mobile, highly likely of the same tracked person. In this case the mobiles are *linked* with a confidence score associated.
best father / son (of a mobile)	the father / son of this mobile with the highest confidence link.
group coherence	measures the likelihood of a real group for a given set of mobiles, at a given frame.
future coherence (of a group)	weighted average of the group coherence over the entire temporal window, where the first frames are given the most importance.

Table 4.3: Specific terminology introduced by Zaidenberg *et al.* [407] regarding to group discovery.

the merging process, computacional efficiency, and formulation of scaled properties to compute the similarity group measure.

The original algorithm considers four main processes for group management and per processed frame, in this order: *update*, *creation*, *termination* and *split/merge*. The *split* process is performed inherently by the technique, not requiring a specific handling of these occurrences, i.e., it is managed by the joint *update* and *creation* processes along the specified temporal window (i.e., 10 frames). The processes also have a time frame associated, in addition to the current frame. The *update*, *creation* and *merge* check a fixed number of frames in the future either for group coherence or finding the common best son, while *termination* checks a fixed number of frames in the past for the outdated status of the maintained mobiles. The *Group Coherence* measure (the larger, less coherent), *GC*, governs some decisions along the aforementioned processes is based on the average distances between mobiles in the same frame, $\tilde{D}_\mu$, and the average over frames of standard deviations of speed, $\tilde{S}_\sigma$, and direction, $\tilde{O}_\sigma$, as follows

$$GC = \omega_1 \cdot \tilde{D}_\mu + \omega_2 \cdot \tilde{S}_\sigma + \omega_3 \cdot \tilde{O}_\sigma \tag{4.1}$$

where (ω_1, ω_2, ω_3) are the weights for each component. Some specific terminology introduced by Zaidenberg *et al.* [407] is also preserved in this thesis, as explained in Table 4.3.

The group discovery in [407] relies on the input provided by the pedestrian detection and tracking module. The obtained measures are projected in the ground-plane and used to compute the similarities between any two objects (mobiles) appearing in a given temporal window, in order to establish possible links among mobiles, with confidence scores associated, in a graph-like representation. However, since our pedestrian tracking is based on the state-of-the-art Boosting tracker [112], which does not provide a graph-based representation, we consider for group discovery just a simple queue of mobiles with one-to-one links (maximum confidence) for the same person and therefore, the best father and best son are always considered the only father and son available for that mobile.

Some improvements are initially performed. Firstly, for efficiency reasons, we only apply the technique on a reduced number of frames, i.e., every second (25 frames in our dataset), without affecting the performance of the group discovery in typical video surveillance environments. This is required to reduce the high computational complexity involved, i.e., going through the temporal window for each frame and for each pair of mobiles. At the same time, we assure a long enough temporal window for more consistent results on group discovery. Secondly, each property of the group coherence measure ($\tilde{D}_\mu$, $\tilde{S}_\sigma$ and $\tilde{O}_\sigma$) is normalized to guarantee stable scaling for any dataset. Thirdly, the weights of these components (ω_1, ω_2 and ω_3) are also normalized to the same end. Moreover, $\tilde{S}_\sigma$ and $\tilde{O}_\sigma$ depend on ancestors and sometimes cannot be computed. In this case, the group coherence measure relies on the only component remaining, $\tilde{D}_\mu$, giving it the full weight (1). Specifically, we set $\omega_i = 0$ if component i is not available and the final measure is normalized by $\sum_{i=1}^{3} \omega_i$. This slightly modified version of the original group discovery algorithm is considered as the baseline in this work.

The preliminary evaluations on our dataset emphasize a frequent limitation of the original technique, namely the unhandled *merge* between an individual and an existing group, to whom he/she belongs. Although by Zaidenberg *et al.* [407] specify a *merge* process between two existent groups, the diversity of group situations from our dataset shows many occurrences where the individuals enter the scene one by one. Even if they enter shortly one after another, any state-of-the-art pedestrian tracking algorithm follows each one once he/she fully enclosed in the view, and this is usually not the same frame for all the persons in a group. Therefore, we conclude that in most video surveillance environments, the initial group *creation* (e.g., usually for the first two individuals) must be followed by such additional *merge* for each individual left that belongs to the group. Otherwise, the remaining participants are never integrated into the group later on, i.e. the incompleteness of the groups is persistent in time, as illustrated in Fig. 4.5(b), namely the third participant of the detected group with green solid line.

Additional experiments show that even if we create a new group for each additional individual, the original *merge* process does not integrate the individual into the group that he/she belongs. Unlike *creation* and *update* processes, the *merge* process is not based on group coherence and this accounts for this shortcoming. Therefore, we address this problem and extend the original technique to handle this situation, i.e., the *creation* process is improved as depicted in Algorithm 2 by adding the lines from 12 to 40. This permits to handle the aggregation of isolated individuals at different locations in the frame.

Our approach for handling the merging individual-group is based on comparing the individual's mobile with the best mobile of each existing group, along the time window. A low value of the group coherence (i.e., below 0.1) indicates a high probability of belonging to that group. Note at line 34 that keeping the best mobile m_{best} is enough to have a reference to the respective group, instead of storing the entire group. Then, at line 44, the actual merge individual-group is performed naturally by the original technique due to the common mobile involved (m_{best}). At this point, the respective mobile m is added to the group and the overall technique starts to track the new member. This version with the merging individual-group improvement is denominated as

Algorithm 2: Creation of groups

input : $groups_{t_c-T-1}$, $mobiles_{t_c-T}$
output: updated $groups_{t_c-T}$

1 $groups_{t_c-T} \longleftarrow \phi$
2 **foreach** $m_i, m_j \in mobiles_{t_c-T}$ **do**
3 $c_f \longleftarrow <>$;
4 **for** $f = t_c - T$ **to** t_c **do**
5 $c_f.append(groupCoherence(m_i, m_j, f))$
6 **end**
7 $c \longleftarrow futureCoherence(c_f)$
8 **if** $c < threshold$ **then**
9 $g_{tmp} \longleftarrow createGroup(m_i, m_j)$ $groups_{t_c-T} \longleftarrow groups_{t_c-T} \cup \{g_{tmp}\}$
10 **end**
11 **end**

`// our contribution, merging individual-group`
12 **for** $m \in mobiles_{t_c-T}$ **do**
`// candidate groups`
13 $g_{cands} \longleftarrow <>$
14 **for** $g \in groups_{t_c-T-1}$ **do**
15 $m_{best} \longleftarrow bestMobile(g)$
`// get the most recent son of best mobile`
16 **while** $bestSon(m_{best}) \neq nullptr$ **do**
17 $m_{best} \longleftarrow bestSon(m_{best})$
18 **end**
`// coherence of` m `and` m_{best} `along time window`
19 $m_{tmp} \longleftarrow m$
20 $g_{tmp} \longleftarrow \phi$
21 $c_f \longleftarrow <>$
22 **for** $f = t_c - T$ **to** t_c **do**
23 $g_{tmp} \longleftarrow g_{tmp} \cup \{m_{tmp}\}$
24 $g_{tmp} \longleftarrow g_{tmp} \cup \{m_{best}\}$
25 $c_f.append(groupCoherence(g_{tmp}, f))$
26 $m_{tmp} \longleftarrow bestSon(m_{tmp})$
27 $m_{best} \longleftarrow bestSon(m_{best})$
28 **if** $m_{tmp} = nullptr \vee m_{best} = nullptr$ **then**
29 break;
30 **end**
31 **end**
32 $c \longleftarrow futureCoherence(c_f)$
33 **if** $c < threshold$ **then**
34 $g_{cands}.append(createGroup(m, m_{best}))$
35 **end**
36 **end**
37 **if** $g_{cands} \neq \{\}$ **then**
`// get the best candidate group for this mobile`
38 $g_{best} \longleftarrow lowestFutureCoherence(g_{cands})$ $groups_{t_c-T} \longleftarrow groups_{t_c-T} \cup \{g_{best}\}$
39 **end**
40 **end**
`// end of contribution`

`// merge groups with common mobiles (see line 9)`
41 **foreach** $g_i, g_j \in groups_{t_c-T-1}$ **do**
42 **for** $m \in gi \cap gj$ **do**
43 $g_{new} \longleftarrow createGroup(g_i, g_j)$
44 $groups_{t_c-T} \longleftarrow groups_{t_c-T} \cup \{g_{new}\}$
45 $groups_{t_c-T} \longleftarrow groups_{t_c-T} \setminus \{g_i, g_j\}$
46 **end**
47 **end**

Group Tracking with Individual-Group Coherence (*GTIGC-V*$_1$).

We additionally improve the group coherence measure by providing a better feature to replace the standard deviation of direction, $\tilde{O}_\sigma$. The new component, $\tilde{C}_{diff}$, measures the average over

(a) (b)

Figure 4.5: Visual comparison between: (a) the analyzed group (ground-truth), indicated by red solid line; and (b) the corresponding detected group (baseline) for the same frame, indicted by green solid line.

frames of the distance difference between each pair of mobiles, which can be interpreted as a measure of sparseness or closeness between the individuals within the group, and is given by Eq. (4.2)

$$\tilde{C}_{diff} = \frac{1}{nF}\sum_{k=1}^{nF}\frac{1}{nP_k}\sum_{i=1}^{n_k-1}\sum_{j=i+1}^{n_k}\sqrt{(m_{i_x}-m_{j_x})^2+(m_{i_y}-m_{j_y})^2} \tag{4.2}$$

where nF is the number of frames containing mobiles that belong to the group, nP_k is the number of pairs of mobiles in frame k, and m_i is the motion vector of mobile i (the vector between the i's best father and i). The original weights of the three components are maintained, i.e., $\omega_1 = 7$ and $\omega_2 = \omega_3 = 5$. The units of $\tilde{D}_\mu$ and $\tilde{C}_{diff}$ are given in meters, $\tilde{S}_\sigma$ is expressed in m/s and $\tilde{O}_\sigma$ in rad. As Section 4.7.4 states, $\tilde{C}_{diff}$ provides a slight performance gain for the overall group discovery. The final proposed group discovery algorithm is so-called *GTIGC-V*$_2$.

4.6.5 Relational Descriptor

We model human behavior in terms of space layout, social environment and nonverbal behavioral interactions. Such social signalling constraints involve attention and position-based cues, which are extracted spatially at each key-point trajectory and temporally at each frame. Indeed, individual and group behaviors can evolve in time, leaving the challenge of simultaneously detecting the transitions between different behaviors and classifying the detected segments. This issue is addressed by working with mini-batches, short enough so that the assumption of constant behavior holds and long enough to enclose sufficient discriminative information.

In terms of I.P.s, the following features are taken:

- *trajectory*, α_{si}, is the angular variation of movement of the individual between consecutive sampling times (in our case, consecutive frames).

- *distance*, d_{io}, expresses the distance between individual position and the object of interest.
- *direction of interest*, β_{gi}, which is the direction of interest, normally called gaze.
- *speed*, v_i, expresses the instantaneous speed of the individual at the sampling time.

In terms of G.B.s, our selection is inspired by the feature-based study of Chamveha *et al.* [140]. Our aim is to simplify feature identification and collection while keeping global discriminative value. This process is expressed by the number of features as well as the number of measurements required to acquire a complete feature. For instance, in [140] they identify four attention-based and five position-based features, and all their measurements, except two, are collected over pair-wise individual relations. In our case, we only consider five features, and only two of them involve pair-wise measurements. Another difference is that in [140] for each feature they consider each single pair-wise relation per sampling step, while in our case we compute a single global contribution for each feature per sampling step.

Considering the G.B.s, the following features are taken at each sampling time:

- *speed*, $\tilde{v}_g$, is the average of the instantaneous velocities of all individuals within a group.
- *distance*, $\tilde{d}_g$, is the average distance between a pair of individuals, considering all the pair-wise relations within a group.
- *speed variance*, $\mathrm{Var}[v_g]$, is the variance of the instantaneous velocities of all individuals within a group.
- *gaze interaction*, $laeo_g$, is a pair-wise relationship and expresses the minimum angle difference between the individual's gaze and the displacement vector between both individual's positions, looking at each other. For each individual, we just considered individuals which fall inside his field of view (knowing the gaze direction and assuming 180° of aperture). This measurement is determined as the mean square error (MSE) of all the differences.
- *profiles*, P_p, reflects the occurrence of I.P.s within a group. In this case, no global measure per sampling step is computed. All profiles contributions are considered individually.

Our descriptor is inspired by Takahashi *et al.* [348]. The features extracted during a pre-defined number of frames are collected and encoded into our fixed-length descriptor to be used in a BoF approach. The key-points along each trajectory are given by

$$\begin{gathered} P_u = \left[p_u^x,\ p_u^y\right] \\ p_u^x = \left[p_u^{x,t_1},\ p_u^{x,t_1+1},\ \ldots,\ p_u^{x,t_2}\right], \quad p_u^y = \left[p_u^{y,t_1},\ p_u^{y,t_1+1},\ \ldots,\ p_u^{y,t_2}\right] \end{gathered} \tag{4.3}$$

where P_u is a set of points in the $\mathcal{T}_u$ trajectory, p_u^x is the set of its x-coordinates, p_u^y is the set of its y-coordinates, and t_1 and t_2 are the starting and ending frames, respectively.

For each feature, the values collected during a mini-batch are encoded into a multi-scale histogram controlled by $R \in \mathbb{N}$, the number of granularity levels. Considering the feature f^0 extracted along the trajectory $\mathcal{T}_u$, its multi-scale representation of size R is given by the vector $\vec{f_u}^0 = \left[H_u^1, H_u^2, ..., H_u^R\right]$, where each entry, H_u^r, is a normalised histogram of 2^{r+1} bins, for each $r = [1, 2, ..., R]$. The final descriptor representation for trajectory $\mathcal{T}_u$ is the concatenation of all the multi-scale feature histograms, and is given by a fixed-length vector

$$\vec{F_u} = \left[\left(\vec{f_u}^0\right), \left(\vec{f_u}^1\right), ..., \left(\vec{f_u}^{N-1}\right)\right] \tag{4.4}$$

4.6.6 BoF Classification

The descriptor is fixed-length to be embedded into a BoF classification approach. The codebook is build by running k-means over a subset of the annotated data, and the obtained centers form the vocabulary to be used on further training and classification processes. We train a multi-class classifier to identify the different I.P.s and G.B.s. The sampling follows a key-point trajectory strategy, where each descriptor is extracted (as explained in Section 4.6.5) over a temporal length, τ, expressed in seconds. Each bag is composed by consecutive descriptors and its length is controlled by Γ. In the case of G.B.s, individual trajectories and gaze orientations within each group are time aligned. The length of the mini-batch, M_i, is given by the number of bags that it might contain, where the shortest case is just one bag, and the largest case is the entire trajectory, I.P., or the group length, G.B.

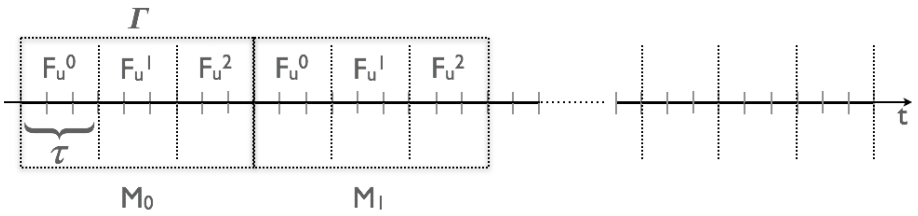


Figure 4.6: Key-point trajectory encoding scheme considering descriptor length and bag length.

The final fixed-length BoF descriptor used for classification can be build at different levels. In this work, we investigate two levels: i) *coarse mini-batch*, where the whole individual trajectory (I.P.) or the entire group set (G.B.) information is used as sample, the fixed-length BoF descriptor is computed from all the bags and the final classification is taken from it; ii) *fine mini-batch*, where the bags are used as individual samples, for each one a fixed-length BoF descriptor is build and a classification is undertaken at the mini-batch level, and the final classification is inferred from a combination of the predicted labels at each mini-batch, either for I.P. or G.B.. On both approaches, the final descriptor vector for each sample is a histogram obtained by nearest cluster counting, which is used as input for an SVM classifier. We adopted a k-fold cross-validation process, maintaining class proportions, to obtain the final classification results for each behavior.

However, in some cases, due to the high class imbalance, we report the classification results under a stratified k-fold cross-validation setting.

We also investigate two components under the classification framework: i) *feature matching*, which is related to the coding step and whose importance relies on a correct cluster histogram matching between the descriptor and the obtained vocabulary; under this component we also studied the impact of the distance measure; ii) *pooling strategy*, which is related to the way the encoded features are summarised to form the final descriptor representation and whose combination defines the discriminative power of the descriptor. For the former one, we normalize the individual feature's histograms and the global descriptor histogram. After that, we compute each histogram matching independently, and combine the distances on the final descriptor by either the average or the max value. For the latter, we change the temporal length of the bag, Γ, and considered two pooling configurations, average and max, for all the descriptors within each bag.

4.6.7 *BACK-F* Relevance Analysis

This technique, so-called *BACKward-Feature selection* (*BACK-F*), is used to inspect feature importance on the final descriptor after the BoF representation. Common feature selection techniques such as *Information Gain* [210] and *Relief-f* [166] cannot be applied directly to histograms because each bin represents a word, which is a combination of several features. In this way, we do a backward procedure starting from the discrete parts of the descriptor (clusters), until the individual feature bins: 1) *cluster ranking*, C_{r_i}, to each cluster is applied the *Relief-f* technique and an importance ranking is obtained; 2) *feature bin ranking*, F_{r_j}, on each cluster the previous step is applied again, resulting on a ranking of bins. Each bin corresponds to an individual feature, described in Section 4.6.5. The final individual feature importance is obtained by

$$F_k = \sum_{i=0}^{C} \sum_{\substack{j=0, \\ l_j=k}}^{B} C_{r_i} \cdot F_{r_j}, \tag{4.5}$$

where C is the number of clusters, B is the number of bins on each cluster, and the condition $l_j = k$ permits to take into consideration the feature bins that correspond to feature's label k. Inspection of this feature selection process is useful to formulate conclusions about the social meaning of each feature.

4.6.8 *KRAV* Relevance Analysis

This technique, denominated by *Kernel-based Relevance Analysis for Video data* (*KRAV*), was implemented in collaboration with the Signal Processing and Recognition Group (SPRG), from Universidad Nacional de Colombia. It is applied directly to the descriptor and its main purpose is to obtain the subset of relevant features, and embed them to improve the final classification. Its formulation is as follows.

Given an input representation space $\boldsymbol{X}\in\mathbb{R}^{N\times P}$ with N samples and P spatio-temporal features and the corresponding social behaviour label vector $\boldsymbol{l}\in Z^N$, we extract the following kernel matrices

$$\boldsymbol{K}_{\boldsymbol{X}} = \kappa_X\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\right) \tag{4.6a}$$

$$\boldsymbol{K}_{\boldsymbol{l}} = \kappa_L\left(l_n, l_{n'}\right) \tag{4.6b}$$

where the former matrix, $\boldsymbol{K}_{\boldsymbol{X}}\in\mathbb{R}^{N\times N}$, holds the pairwise similarity between row-vector samples $\boldsymbol{x}_n, \boldsymbol{x}_{n'}\in\boldsymbol{X}$, and $n, n'\in[1,2,\ldots,N]$. Grounded on its general approximating ability and mathematical tractability, we select the Gaussian kernel defined as below

$$\kappa_X\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \sigma\right) = \exp\left(-\mathrm{d}^2\left(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\right)/2\sigma^2\right) \tag{4.7}$$

where $\mathrm{d}(\cdot,\cdot):\mathbb{R}^P\times\mathbb{R}^P\mapsto\mathbb{R}^+$ is a distance operator and $\sigma\in\mathbb{R}^+$ is the kernel bandwidth and the latter kernel matrix, $\boldsymbol{K}_{\boldsymbol{l}}\in\mathbb{R}^{N\times N}$, holds the pairwise similarity between labels $l_n, l_{n'}\in\boldsymbol{l}$. Namely, we set a positive definite kernel for the label kernel κ_L that measures the pairwise similarity of l_n, $l_{n'}$ labels as below

$$\kappa_L\left(l_n, l_{n'}\right) = \delta_{l_n l_{n'}} \tag{4.8}$$

where the delta function is $\delta_{l_n l_{n'}}=1$ if the samples have the same label $l_n=l_{n'}$, otherwise, $\delta_{l_n l_{n'}}=0$.

To assess the joint information between the spatio-temporal features and the corresponding social behaviour labels, we evaluate how well the estimated kernel functions, κ_X and κ_L, align to each other. To this end, the commonly-known centered kernel alignment (CKA) that measures the similarity between a couple of characterizing kernel functions [113] is applied. In particular, we employ the normalized inner product of both kernel functions to estimate the dependence between jointly sampled data as follows [47]

$$\rho\left(\bar{\boldsymbol{K}}_{\boldsymbol{X}}, \bar{\boldsymbol{K}}_{\boldsymbol{l}}\right) = \frac{\langle\bar{\boldsymbol{K}}_{\boldsymbol{X}}, \bar{\boldsymbol{K}}_{\boldsymbol{l}}\rangle_{\mathrm{F}}}{\sqrt{\langle\bar{\boldsymbol{K}}_{\boldsymbol{X}}, \bar{\boldsymbol{K}}_{\boldsymbol{X}}\rangle_{\mathrm{F}}\langle\bar{\boldsymbol{K}}_{\boldsymbol{l}}, \bar{\boldsymbol{K}}_{\boldsymbol{l}}\rangle_{\mathrm{F}}}} \tag{4.9}$$

where notation $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ stands for the Frobenius-based matrix inner product that is defined for two matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ as $\langle\boldsymbol{U},\boldsymbol{V}\rangle_{\mathrm{F}}=tr(\boldsymbol{U}^\top\boldsymbol{V})$, $\bar{\boldsymbol{K}}$ is the centered version of the kernel matrix $\boldsymbol{K}$ calculated as $\bar{\boldsymbol{K}}=\tilde{\boldsymbol{I}}\boldsymbol{K}\tilde{\boldsymbol{I}}$, $\tilde{\boldsymbol{I}}=\boldsymbol{I}-\frac{1}{N}\boldsymbol{1}^\top\boldsymbol{1}$ is the empirical centering matrix, $\boldsymbol{I}\in\mathbb{R}^{N\times N}$ is the identity matrix, and $\boldsymbol{1}\in\mathbb{R}^N$ is the all-ones vector.

Besides, we rely on the Mahalanobis distance to carry out the pairwise comparison between samples $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ based on the Gaussian kernel κ_X. Therefore, the distance function in Eq. (4.7) is fixed as follows

$$\mathrm{d}^2_{\boldsymbol{A}}(\boldsymbol{x}_n, \boldsymbol{x}_{n'})=(\boldsymbol{x}_n-\boldsymbol{x}_{n'})\,\boldsymbol{A}\boldsymbol{A}^\top\,(\boldsymbol{x}_n-\boldsymbol{x}_{n'})^\top \tag{4.10}$$

where matrix $\boldsymbol{A}\in\mathbb{R}^{P\times M}$ holds the linear projection $\boldsymbol{y}_n=\boldsymbol{x}_n\boldsymbol{A}$, with $\boldsymbol{y}_n\in\mathbb{R}^M$, $M\leq P$. The Euclidean distance of the M-dimensional feature space $d^2(\boldsymbol{y}_n, \boldsymbol{y}_{n'})$ is equivalent to a Mahalanobis distance

in the original space, with the inverse covariance matrix $\boldsymbol{A}\boldsymbol{A}^\top$. Namely, $\boldsymbol{\Sigma}_X^{-1}=\boldsymbol{A}\boldsymbol{A}^\top$, being $\boldsymbol{\Sigma}_X\in\mathbb{R}^{P\times P}$ the covariance matrix of $\boldsymbol{X}$.

To compute the projection matrix $\boldsymbol{A}$, the formulation of CKA-based function in Eq. (4.9) can be integrated into the following kernel-based learner

$$\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \log\left(\rho\left(\bar{\boldsymbol{K}}_{\boldsymbol{X}}, \bar{\boldsymbol{K}}_l; \boldsymbol{A}\right)\right) \tag{4.11}$$

where the logarithm function is used for mathematical convenience. The explicit objective function of the empirical CKA in Eq. (4.9) yields [47]

$$\hat{\rho}\left(\boldsymbol{K}_{\boldsymbol{X}}, \boldsymbol{K}_l\right) = \log\left(tr\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_l\tilde{\boldsymbol{I}}\right)\right) - \tfrac{1}{2}\log\left(tr\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\right)\right) + \rho_0 \tag{4.12}$$

where $\rho_0\in\mathbb{R}$ is a constant that we assume does not depend on $\boldsymbol{A}$.

Consequently, the optimizing approach in Eq. (4.11), besides learning the optimal projection matrix $\hat{\boldsymbol{A}}$, also demands tuning of the Gaussian kernel bandwidth σ. To deal with the joint parameter estimation, we propose to optimize iteratively one variable at a time while the other variable is fixed.

When maximizing parameter $\boldsymbol{A}$ with fixed σ, the gradient function of the objective function in Eq. (4.12) results in the form

$$\nabla_{\boldsymbol{A}}\left(\hat{\rho}\left(\boldsymbol{K}_{\boldsymbol{X}}, \boldsymbol{K}_l\right)\right) = -4\boldsymbol{X}^\top\left(\left(\boldsymbol{G}\circ\boldsymbol{K}_{\boldsymbol{X}}\left(\boldsymbol{A},\sigma\right)\right) - \mathrm{diag}\left(\mathbf{1}^\top\left(\boldsymbol{G}\circ\boldsymbol{K}_{\boldsymbol{X}}\left(\boldsymbol{A},\sigma\right)\right)\right)\right)\boldsymbol{X}\boldsymbol{A} \tag{4.13}$$

where notations $\mathrm{diag}(\cdot)$ and $\circ$ denote the diagonal operator and the Hadamard product respectively. $\boldsymbol{G}\in\mathbb{R}^{N\times N}$ is the gradient of the objective function with respect to $\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)$, calculated as follows

$$\boldsymbol{G} = \nabla_{\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)}\left(\hat{\rho}\left(\boldsymbol{K}_{\boldsymbol{X}}, \boldsymbol{K}_l\right)\right) = \frac{\tilde{\boldsymbol{I}}\boldsymbol{K}_l\tilde{\boldsymbol{I}}}{tr\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_l\tilde{\boldsymbol{I}}\right)} - \frac{\tilde{\boldsymbol{I}}\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}}{tr\left(\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{A},\sigma)\tilde{\boldsymbol{I}}\right)} \tag{4.14}$$

For updating the estimation of $\boldsymbol{A}$, we use the standard stochastic gradient descent update rule, provided the initial guess $\boldsymbol{A}^o$, as follows

$$\boldsymbol{A}^{t+1} = \boldsymbol{A}^t - \mu_A^t\nabla_{\boldsymbol{A}^t}\left(\hat{\rho}\left(\boldsymbol{K}_{\boldsymbol{X}}, \boldsymbol{K}_l\right)\right) \tag{4.15}$$

where $\mu_A^t\in\mathbb{R}^+$ is the step size of the learning rule. $\boldsymbol{A}^t$ and σ^t are the samples we use at the time step t.

As a result, the estimated projection matrix $\widehat{\boldsymbol{A}}$ is differently affected by each input feature. Consequently, we quantify the contribution of each input feature for building $\widehat{\boldsymbol{A}}$ by introducing

the relevance vector index, $\boldsymbol{\varrho} \in \mathbb{R}^P$ with elements computed as below

$$\rho_p = \sum_{m=1}^{M} |a_{pm}|; \forall p \in P, \quad a_{pm} \in \widehat{\boldsymbol{A}} \tag{4.16}$$

The main rationale behind the proposed relevance index is that the features with bigger variability should favor a better kernel adjustment. Thus, the larger ρ_p the better the input features. Furthermore, we rank the original feature set in terms of ρ_p to perform the selection of the most discriminant spatio-temporal features. In this way, we choose M_S features, $M_S \leq P$, where their value of $\boldsymbol{\varrho}_p$ exceed a threshold. This is set to avoid the computation of all the histogram descriptors, resulting $\boldsymbol{X}_S \in \mathbb{R}^{N \times M_S}$.

Nonetheless, we further enhance the class inter-separability through the following embedding matrix $\boldsymbol{X}_E = \boldsymbol{X}_S \widetilde{\boldsymbol{A}}$, where $\boldsymbol{X}_E \in \mathbb{R}^{N \times M_E}$ and $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{M_S \times M_E}$ is another projection matrix computed from $\boldsymbol{X}_S$ instead of $\boldsymbol{X}$ following the optimization problem in Eq. (4.11), being $M_S \geq M_E$. The value of M_E is selected by variance-based relevance criteria where 90% of variance of $\boldsymbol{X}_S$ is retained.

4.7 Validation

4.7.1 Camera Calibration and Ground-Plane Projection

During camera calibration, our automatic procedure permits to acquire image points with significant variations in pose (rotation) and depth (translation). This improves the fitting of the camera model, since bigger set of parameters, namely angles and positions, are given instead of redundant data that probably add noise. The final re-projection error is expressed as the average of the root mean square of the difference between the image points used to compute the object points and the recalculated image points from the obtained camera model. The reported error is very low (0.117 pixels).

For ground-plane projection, we use the points identified in Fig. 4.1(b) to estimate the image-to-plane projection error. Taking in consideration the rectangle pattern between subsequent pair of segments, both in parallel and perpendicular directions, we compute the absolute errors of collinear segments in real world measure. The results vary between $[0.34\%, 4.91\%]$.

4.7.2 Pedestrian Tracking

Considering the selected tracking algorithms in Section 4.6.2, namely Boosting and Boosting-Improved, we obtain the results for the measures MOTP and MOTA that show the best performance of the Boosting-Improved approach, where a significant improvement is achieved in both measures (see Table 4.4). The analysis of Fig. 4.7 corroborates the previous conclusion, since the density of the trajectories extracted from Boosting-Improved algorithm are closer to the density of the manual trajectories, than the trajectories extracted from the original Boosting, which exhibit problems related to the temporal continuity of the tracking process. However, we also verify that

the Boosting-Improved algorithm maintains some errors from spatial variability. This probably comes from the way the features are weighted in the multi-assignment process. Fig. 4.8 illustrates some cases where Boosting-Improved clearly presents better tracking results, especially in terms of continuity, than the original Boosting tracker.

	Boosting-Improved	Boosting
MOTP	**10.1**	12.2
MOTA	**79.9**	71.3

Table 4.4: Tracking performance (%) of Boosting-Improved and Boosting given by the MOTP and MOTA metrics. For MOTP, the lower is the better, while in MOTA, the higher is the better.

(a) (b) (c)

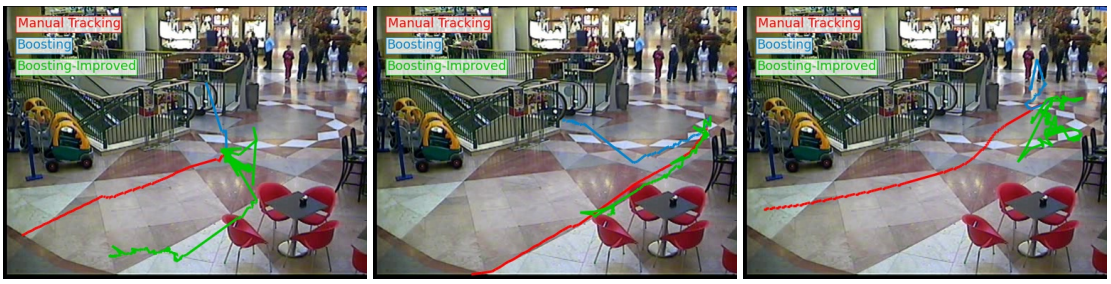
Figure 4.7: Subsample ($\approx 25\%$) of the trajectories obtained from: (a) manual annotation, (b) Boosting-Improved algorithm, (c) Boosting algorithm.

(a) (b) (c)

Figure 4.8: Examples of individual trajectories where Boosting-Improved overperform the Boosting algorithm.

Fig. 4.9 illustrates the comparison of both algorithms results with the manual annotation for individual trajectories, where we observe tracking failures. In fact, despite the recent advances, the state-of-the-art algorithms still underperform in cluttered environments that present many occlusions. Additionally, the trackers in consideration are single-track, therefore they do not take into account the multi-person coexistence and interactions in order to jointly improve the final tracking. We should highlight that none pre-processing technique such as filtering, background subtraction,

among others, and post-processing technique like non-maximum suppression or scene knowledge is used to improve the performance of the chosen tracking algorithms. It is also important to mention the importance of this low-level processing module, which may drastically impair the classification results if its computation is not reliable enough (see Section 4.7.5.4 for the evaluation of its impact).



(a) (b) (c)

Figure 4.9: Example of tracking failures from Boosting-Improved and Boosting algorithms.

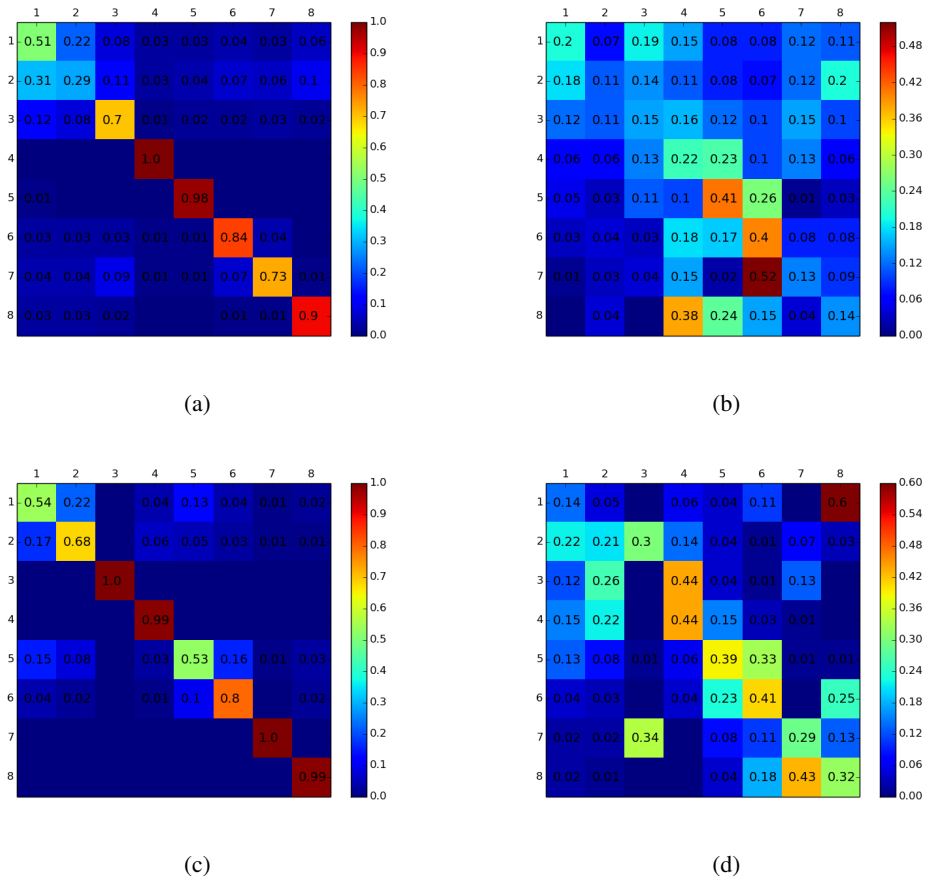
4.7.3 Gaze Estimation

The gaze estimation performance consider two tracking data, from Boosting-Improved and from ground-truth. The evaluation is assessed by two different methods: i) from Chamveha *et al.* [61], which considers the auto-determined walking directions as the ground-truth labels of the persons' head orientations; ii) from ground-truth (GT), using our annotated data. The results are shown in Table 4.5 and reported on the three standard metrics, namely precision (P), recall (R) and accuracy (A). Fig. 4.10 shows the corresponding confusion matrixes, and Fig. 4.11 illustrates the number of representative images per gaze orientation index automatically generated.

Tracking	Metric	Gaze estimation (evaluation from [61])	Gaze estimation (comparing with GT)
Boosting-Improved	P	49.0	6.4
	R	50.8	20.4
	A	87.3	84.6
GT	P	59.5	7.0
	R	54.3	13.9
	A	89.7	84.5

Table 4.5: Evaluation of the gaze estimation performance (%) for two tracking data (Boosting-Improved and GT) and for two evaluation methods (Chamveha et al. [61] and comparing with GT).

As expected, the evaluation method from Chamveha *et al.* [61] reports significantly better results, as depicted in the third column of Table 4.5. However, the generated samples are additionally tested with our manual annotation and the expected deterioration of the results is shown in the fourth column. Finally, Table 4.5 also shows slightly better results for manual annotation as



(a) (b)

(c) (d)

Figure 4.10: Confusion matrixes for gaze estimation: for two tracking data (Boosting-Improved in (a), (b) and GT in (c), (d)), and for two evaluation methods (Chamveha et al. [61] in (a), (c) and comparing with GT in (b), (d)).

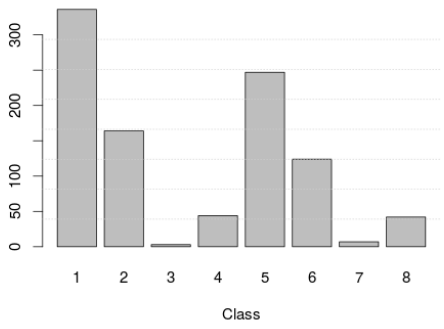


Figure 4.11: The obtained number of representative images per gaze orientation index.

tracking data, since the tracking loss usually causes less representative images for head orientation and even false positives in some cases.

The confusion matrixes from Fig. 4.10 correspond to the same four situations from the Table 4.5, respectively, and it can be observed the same difference in the results obtained with the two evaluation methods, when comparing 4.10(a), 4.10(c) with 4.10(b), 4.10(d). In particular, the nearly perfect classification of the labels 3, 4, 7, 8 from Fig. 4.10(c) is caused by the very few samples for those walking directions, as illustrated in Fig. 4.11. The work in [61] uses oversampling to handle imbalanced data, but it shows no effect in our dataset where, due to environment constrains, the constant flow of pedestrians is restricted to a diagonal path where the 1, 2, 5, 6 walking directions are very frequent (see Fig. 4.11), causing a severe imbalance of the samples.

4.7.4 Group Discovery

We aim to extensively evaluate our proposal for group discovery in typical video surveillance environments comprising diverse, unconstrained and unscripted group dynamics in order to demonstrate the robustness of the technique in the real world. As previously introduced in Section 4.4, the dataset provides nearly one hour of recording from a shopping mall, with 750 persons, 211 groups, and group members ranging from 2 and 9. Additionally, it provides diverse group dynamics (creation, split, merge, termination) that motivates the focus of our evaluation on this dataset.

The results of group discovery are presented in Tables 4.6 and 4.7. Our approach for group discovery comprises two versions, corresponding to the contributions discussed in Section 4.6.4: i) *GTIGC-V*$_1$ includes only the merging individual-group in addition to the baseline; ii) *GTIGC-V*$_2$ includes both merging and $\tilde{C}_{diff}$ measure, as alternative to $\tilde{O}_{\sigma}$.

The evaluation starts with a group identification process before applying the actual metrics for performance comparison. This step aims to determine, for each group from ground truth, its best match among the detected groups. A score is computed when comparing two groups, as the total number of mobile matches between the two groups, divided by the maximum number of mobiles of both groups. The detected group with the highest score is finally selected for performance evaluation. This group is most likely the aimed one. An example is illustrated in Fig. 4.5.

Firstly, we compare our technique with the baseline using our dataset. We also aim to evaluate the performance impact of automatic tracking in the group discovery algorithm. Therefore, Table 4.6 shows results with and without automatic tracking (and also gaze) information. For this experiment, the evaluation methodology of Hung and Ben [138] is employed. The precision (P), recall (R) and F_1 measures for group discovery are computed as follows: any overlapping between the labeled and detected groups is given a score depending on whether the detection is a true positive, false positive or false negative, i.e., whether that person is in the labeled group in that particular frame; the scores are accumulated to get the precision, recall and F_1 measures of that particular group; at the end we compute the average of each of these measures over all the groups.

As expected, the recall improved by properly integrating the remaining participants into their groups, (see Fig. 4.12), using *GTIGC-V*$_1$. In some cases, this process also integrates people in wrong groups (see Fig. 4.13) and those false positives account for lower precision. As expected, accuracy and F_1 maintain similar levels given these two opposite variations. However, through a qualitative analysis, we verify that our contribution on merging remaining individuals to their

Technique	PDT	Gaze	P	R	A	F_1
Baseline	M	M	**86.5**	73.6	69.4	81.7
GTIGC-V$_1$	M	M	82.3	**77.2**	70.1	81.6
GTIGC-V$_2$	M	M	84.2	76.5	**70.7**	**82.6**
GTIGC-V$_2$	A	A	62.4	76.3	51.0	64.4

Table 4.6: Results on group discovery (%) w.r.t. the baseline for our dataset, for Manual (M) annotation and Automatic (A) *Pedestrian Detection and Tracking* (PDT) and gaze. The last columns report the evaluation measures, namely precision (P), recall (R) and F_1.

groups solves the vast majority of such cases, achieving in a semantical sense the completeness of the groups. Therefore, we consider this approach as the first step towards a robust group discovery in most video surveillance environments, and the remaining problem of the false positives must be tackled in our future work. Additionally, *GTIGC-V*$_2$ shows slightly better results over both baseline and *GTIGC-V*$_1$, emphasizing the benefits of the proposed $\tilde{C}_{diff}$ measure. The impairment caused by the automatic tracking in the group discovery performance is also verified.

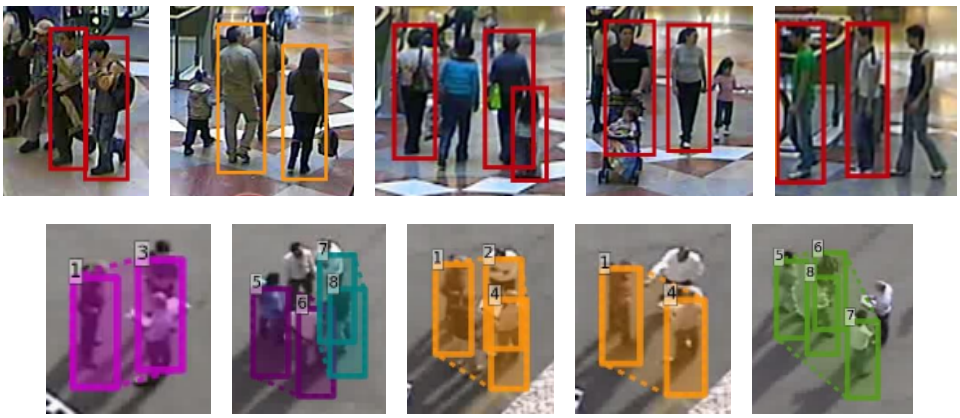


Figure 4.12: Examples of incomplete groups, from our dataset (first row) and FM (second row), that were successfully corrected by the *GTIGC-V*$_1$ version w.r.t. the baseline. Each group is identified with the same colour.

Secondly, we compare our technique for group discovery with Bazzani *et al.* [30] using their Friends-Meet (FM) dataset. The FM dataset comprises a synthetic and a real set. The real set, used in this work, contains groups of people recorded in an outdoor area and mainly in F-formations and with, split and merge events (person-to-group and group-to-group), and queue formation. It provides simple group dynamics in a static scene from a bird's-eye view with no occlusions, constant relative distances among members of the same group, constant person speeds, etc. The recording contains 10698 frames, 960×720 resolution at 30 fps, organized in 15 sequences of lengths between 30 and 90 seconds, containing between 3 and 11 individuals acting with a script. FM provides manually annotated people and groups in each frame, i.e., for each individual, the position and velocity in the image, ground floor position and velocity, personal identifier and group identifier that he/she belongs to.

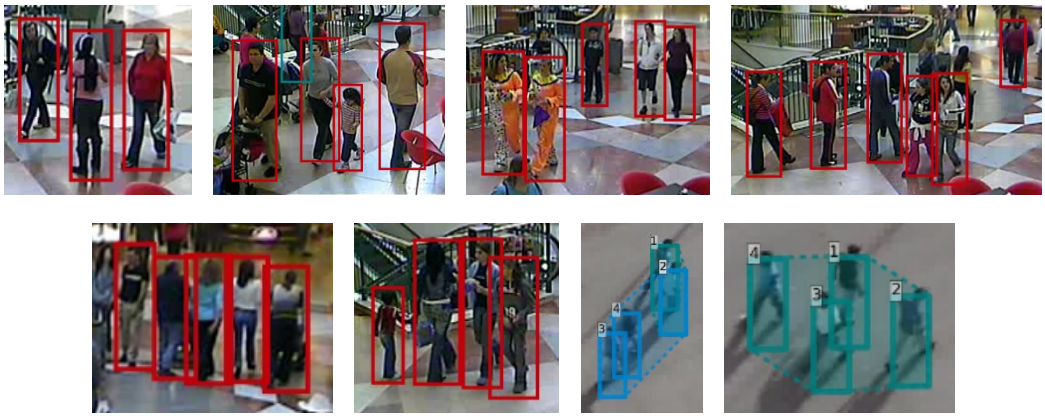


Figure 4.13: Examples of false groups, from our dataset (the first 6) and FM (the last 2), as outcome of the *GTIGC-V*$_1$ version w.r.t. the baseline. Each group is identified with the same colour.

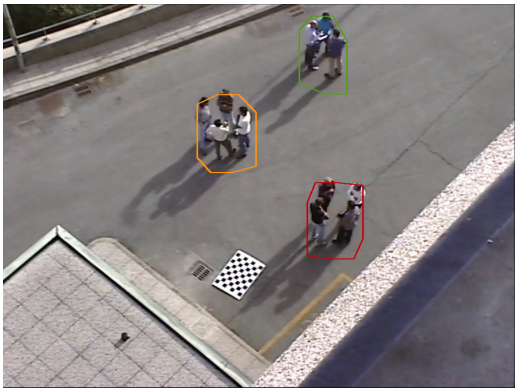


Figure 4.14: Example of detected groups in the Friends-Meet dataset [30] and the convex hull evaluation approach.

For this experiment, the evaluation is modified to use the convex hull around the members of the group, instead of considering individual bounding-boxes for each person, in order to accomplish a fair comparison with the results of the state-of-the-art methods stated by Bazzani *et al.* [30]. Fig. 4.14 illustrates an example of the detected groups. The metrics of this experiment enclose the ones presented in Bazzani *et al.* [30], namely: a) 1 - False Positive (FP), b) 1 - False Negative (FN), c) Group Detection Success Rate (GDSR), d) Multi-Object Tracking Precision (MOTP) and e) Multi-Object Tracking Accuracy (MOTA). All these measures concern the evaluation of the correct detection of groups over the time, namely formation, life-span, and dispersion.

Bazzani *et al.* [30] adopt the metrics a, b from Smith *et al.* [336] and d, e from Bernardin and Stiefelhagen [34]. Table 4.7 shows two variants of the proposed algorithm by [30]: *DEEPER-JIGT*, which uses manually annotated data at the beginning of the tracker and each time the target is lost; *DEEPER-JIGT.2*, that always uses manual information. Our baseline and *GTIGC-V*$_2$ use automatic tracking information from the Boosting-Improved algorithm. As shown in Table 4.7, our technique improved the MOTP w.r.t. the baseline due to the correct merging of the groups. In

this dataset, most of the group formations that are persistent for longer periods of time comprise four people (i.e., static, circular arrangement), and are initially formed by separate small groups of two people that enter the scene together. While standing side-by-side, the baseline detects the four persons as two separate groups. Therefore, correcting this shortcoming in *GTIGC-V_2*, a higher precision is obtained. The lower the MOTP, the better. *DEEPER-JIGT* achieved slightly higher GDSR than our technique. However, as introduced by Bazanni *et al.* [30], the GDSR assumes that a group is correct if at least 60.0% of its members are detected. Therefore, given the few number of members in groups and small variation in this number, we argue that more complex scenarios have to be considered in order to produce more relevant results.

Technique	1-FP (%)	1-FN (%)	GDSR (%)	MOTP (m)	MOTA (%)
DEEPER-JIGT.2 [30]	97.1	**93.8**	**88.5**	**0.64**	71.7
DEEPER-JIGT [30]	95.6	91.1	86.1	0.80	67.6
Baseline	98.7	**94.9**	81.8	0.09	**92.5**
GTIGC-V_2	**99.2**	85.9	84.1	0.04	84.4

Table 4.7: Results on group discovery w.r.t. [30] (DEEPER-JIGT, DEEPER-JIGT.2) for the Friends-Meet dataset and convex hull evaluation approach. For MOTP, the lower is the better, while in MOTA, the higher is the better.

We also conducted a qualitative evaluation, where we have notice that their annotation take into consideration principles different from ours, for instance, some individuals that walk in the direction of each other are instantaneously aggregated into the group, even when they are far away. We believe that this is the main cause for such impairment, apart from the tracking failures since the FM dataset is very challenging for tracking due to occlusions and low resolution.

4.7.5 BoF Classification Settings

To analyse our descriptor performance, we compare the classification results with a baseline descriptor, which is composed by the same type of features enumerated in Section 4.6.5, but instead of considering a multiscale histogram from the features, it simply considers the mean, μ, and standard deviation, σ, of each feature, except for the P_p feature. For the case of G.B.s, we add a state-of-the-art descriptor, referred here as Chamveha, that uses our multiscale descriptor formulation but includes the features presented in Chamveha *et al.* [140]. Under our experiments, $R = 3$ showed a good trade-off between accuracy and dimensionality length, which leads to a 112 and 116 dimensional descriptor vector for I.P.s and G.B.s, respectively. Each feature included in the descriptors are normalized; for instance the ones that consider distances are normalized by the diagonal of the detected ground-plane, speeds by the normal walking speed ($\approx$ 6.0 km/h), and angles by 2π.

For an exhaustive classification evaluation, we adopt a two-fold cross-validation repeated over 10 random iterations. In order to obtain fair results we keep classes' proportions from the original dataset for each fold. Since the imbalance among the classes is very high (see Table 4.1), we randomly replicate the samples by a percentage, ρ, in the training fold, while maintaining fixed

the number of samples of the most representative class and increasing proportionally the remaining classes. The evaluation considers three standard parameters: accuracy (A), recall (R), and precision (P); and sometimes, when relevant, the F1-score (F_1) is also presented.

Using the trajectories and gazes manually annotated, we run experiments over different parameter settings and compared results over an overall F_1. The experiments consider several classifiers, replication percentage of the training set, variations on descriptor temporal length τ, bag size Γ, and number of clusters in BoF, K. Several classifiers are tested such as Multi-Layer Perceptrons (MLP), random trees, gentle AdaBoost, normal Bayes, and Support Vector Machine (SVM), among others. For sake of simplicity, we only report in this thesis the most significative results to support our conclusions. We observe that the SVM classifier presents the best performance, therefore a deeper analysis is conducted over different kernels, namely linear, polynomial, Radial-Basis Function (RBF) and intersection, verifying the best results for the intersection kernel. Table 4.8 summarizes the tested parameters and their best values verified empirically. The results presented from this Section until Section 4.7.5.5 are evaluated under this BoF classification setting.

ρ (%)	K	τ (s)		Γ	
		I.P.	G.B.	I.P.	G.B.
15	70	5	1	1	7

Table 4.8: Empirical values for some parameters of ***VISOBI*** using BoF representation.

4.7.5.1 Feature Strategy and Matching Distance

This component is related to the coding step of the descriptor into the BoF representation, where the best cluster histogram matching between the descriptor and the obtained vocabulary is desired. To analyse this component, we compare two histogram matching techniques, namely the average and the max, where the distances between the individual feature's histograms of the final descriptor are computed and the final distance is considered to be the average or the max of them, respectively. In this way, the distances from all clusters are stored and a decision is made taking one of both techniques. For evaluation, we consider the F_1 measure of both matching strategies fixing the pooling scheme.

Inspecting Table 4.9, we verify that in general the average matching presents better results than the max. This difference is lower when classifying the I.P.s, specially when the max is taken, but a drastic performance drop is obtained while classifying the G.B.s. This can be explained by several factors: i) more variability, since each group is represented by a global behavior that depends on the number of individuals within it and their profile's variance; ii) more separability among G.B.s, by definition two of the classes are highly distinguishable, namely *Exp.* and *CHAT.*; iii) similar features contribution for the descriptor; iv) lower correlation among features, which is pretty well represented looking at the values of the correlation distance in Table 4.9.

Concerning the matching distance, it is obvious that the correlation distance is the worst, while the remaining present similar performance. Since the average matching technique presents better

Matching	Average						Maximum					
Pooling	Avg.			Max.			Avg.			Max.		
Labels	I.P.s	G.B.s	All	I.P.s	G.B.s	All	I.P.s	G.B.s	All	I.P.s	G.B.s	All
Intersection	**55.6**	**58.0**	**56.8**	42.7	53.2	48.0	38.1	36.3	37.2	41.4	32.3	36.9
Euclidean	41.7	58.0	49.9	44.3	53.1	48.7	38.4	36.4	37.4	36.2	11.3	23.8
Correlation	40.1	52.4	46.3	42.3	51.7	47.0	39.1	9.6	24.4	37.1	10.8	24.0
Bhattacharyya	43.3	50.7	47.0	44.0	54.2	49.1	38.4	38.0	38.2	39.1	36.5	37.8

Table 4.9: Mean F1-score (%) of I.P.s, G.B.s and overall for the combination of histogram matching, distance measure and pooling configurations, using our descriptor.

results for all the classes, we select the intersection distance, which also reveals the best classification. The combination of the histogram intersection measure with the intersection kernel SVM corroborates that the combination of both generate better visual codebooks under unsupervised learning [381].

4.7.5.2 Pooling Strategy

The goal of the pooling strategy is to achieve invariance over possible transformations, provide compact representations and achieve higher performance removing irrelevant information. Indeed, the pooling strategy could modify the BoF representation. In this way, we investigate if the temporal length of bags, Γ, and their mode of aggregation affect the final classification performance.

Overall evaluation confirms that average pooling technique performs better than the max pooling technique, specially when classifying the G.B.s, which makes sense since more information is collected by bag. The difference between both is small. This leads us to conclude that since each descriptor is extracted by key-point trajectory sampling, all the sampling points are relevant for the final representation of the trajectory.

4.7.5.3 Descriptor Performance

In this section, we simulate two anomalous behaviors that might affect the descriptor performance: i) *tracking loss*, where some trajectories' segments are removed; ii) *noise variation*, where different degrees of noise in terms of the variation of σ are injected into the trajectories and gazes. Both conditions try to simulate tracking and gaze estimation loss and errors. For the training test we use samples without any kind of perturbation.

From Fig. 4.15(a) and Fig. 4.15(b) we verify a slight performance drop with the increase of noise. However, what is important to retain is that the fluctuation of performance is higher on I.P.s than on G.B.s, which means that G.B.s are less sensitive to *tracking loss* and that our descriptor can characterize small temporal segments with a similar performance than the whole set of segments that constitute the I.P. or G.B. Both statements confirm the evidence that will be stated in next Section 4.7.5.5, that shows that the *fine mini-batch* approach presents a performance similar to the

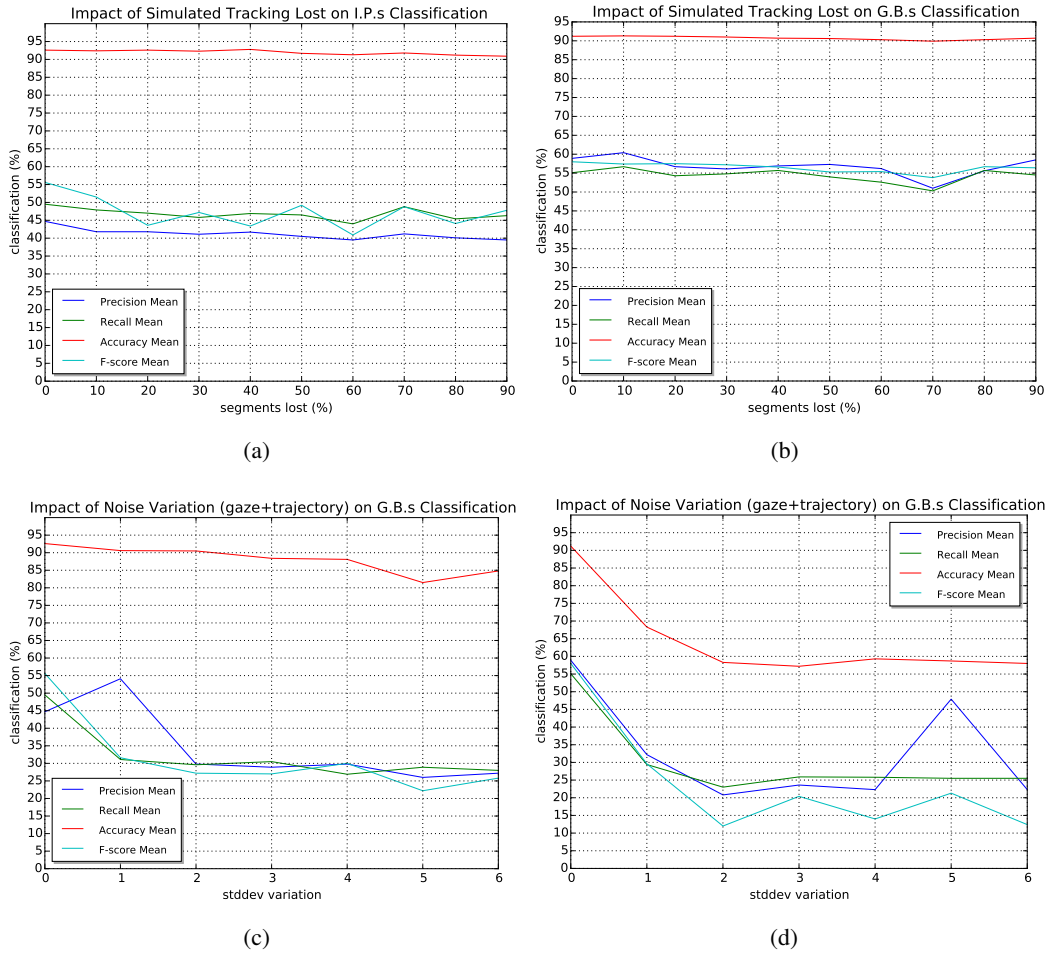


Figure 4.15: Simulation of the impact of *tracking loss* (for I.P. in (a) and G.B. in (b)) and *noise variation* in gaze and trajectory (for I.P. in (c) and G.B. in (d)) in our descriptor considering the classification results.

coarse mini-batch approach for G.B.s and a slightly worse for I.P.s. We should also highlight that the simulation of segments loss is done at the level of the bag and not at the level of the key-points (tracking level), therefore despite the remotion of some segments, the remaining ones keep the temporal structure.

The *noise variation* in gaze and trajectory also affects the performance. Fig. 4.15(c) and Fig. 4.15(d) show a decreasing function with an initial steepest drop. As expected, the negative impact on G.B.s is higher than the I.P.s due to the relational features involved among individuals of the same group. The decrease in performance could have been even higher if there is not the compensation effect of the average velocity and average distance. As Fig. 4.21(b) shows, those two features contribute the most to the recognition of the G.B.s. The respective average attenuates the fluctuations in velocity and distance, caused by noise variation in trajectory. In fact, such smoothing could be the reason for some oscillations in the performance curve even with the increase of noise.

4.7.5.4 Impact of Automatic Pedestrian Tracking and Gaze on Classification

In this experiment, we consider the *coarse mini-batch* approach. Table 4.10 presents the impact of the Boosting-Improved algorithm, for the extraction of trajectories, on the overall classification results when compared with the manual annotation, for I.P.s and G.B.s (see rows AT-MG and MT-MG, respectively). As expected, the automatic tracking causes a deterioration of the results. Inspecting the second and third rows, we verify that most of the performance drop from the automatic feature extraction is due to tracking failures (see Fig. 4.9).

	E.I.				B.I.				U.I.				CHAT.				Avg.			
	P	R	A	F_1	P	R	A	F_1	P	R	A	F_1	P	R	A	F_1	P	R	A	F_1
MT-MG	90.8	93.9	87.9	92.3	52.7	56.8	91.9	54.0	56.6	42.7	89.6	46.7	35.6	27.0	95.6	38.9	58.9	55.1	91.2	58.0
MT-AG	90.0	91.7	85.8	90.8	47.9	41.4	91.3	42.8	42.2	41.0	87.7	43.0	27.0	29.0	94.1	29.4	51.8	50.8	89.7	51.5
AT-MG	78.8	82.2	69.3	80.4	14.2	13.6	85.6	18.3	12.4	10.3	81.4	13.9	15.6	8.0	94.8	25.6	30.3	28.5	82.8	34.5
AT-AG	79.2	78.4	67.6	78.7	10.2	11.8	84.1	14.3	14.7	16.0	80.9	15.5	11.2	8.0	93.6	25.5	28.9	28.5	81.6	33.5
	Dist.				Exp.				Dis.				Int.				Avg.			
MT-MG	28.9	44.1	89.4	34.3	95.3	89.8	86.8	92.4	11.6	10.0	98.5	48.8	43.2	54.2	95.7	46.7	44.7	49.5	92.6	55.6
MT-AG	32.4	40.2	90.8	35.4	95.4	90.7	87.8	93.0	12.3	30.0	97.6	25.8	54.6	64.7	96.5	57.0	48.7	56.4	93.2	52.8
AT-MG	17.6	30.1	86.4	21.8	92.4	78.8	75.1	85.0	9.7	22.5	98.2	27.5	7.8	23.6	87.3	11.6	31.8	38.7	86.7	36.5
AT-AG	13.6	23.2	85.3	16.9	92.3	79.0	75.3	85.1	24.6	42.5	89.7	40.5	7.3	21.7	87.4	10.8	34.5	41.6	86.7	38.4

Table 4.10: Classification results (%) for all I.P.s and G.B.s. considering our descriptor and combinations of Manual (M) and Automatic (A) feature extraction processes for Tracking (T) and Gaze (G).

In general, the negative impact is higher on G.B.s than I.P.s, which is obvious, since one of the features used to identify a G.B is the inferred label of the I.P. that can be affected directly by the automatic tracking. We also verifiy that *B.I.* and *U.I.* are the most impaired, which is also expected since their behavior is highly dependent on their trajectories, while *CHAT.* is mostly affected by detection, since it represents Free-standing Conversational Groups and *E.I.* is the predominant class with a large number of samples. In terms of I.P.s, the *Int.* is the most compromised, the *Exp.* is the less affected since it is the most representative class, and the *Dist.* and *Dis.* reveal the largest significative drop since, by definition, they are the most dependent on the trajectory behavior, specially the *Dis.* which in fact shows the worst result with automatic tracking. Indeed, all the results are affected by the eventual loss of tracks and their random movement later on (see Fig. 4.9).

Table 4.10 also presents the impact of gaze estimation, for the extraction of head directions, on the overall classification results w.r.t. the manual annotation, for I.P.s and G.B.s (see rows MT-AG and MT-MG, respectively). We verify that all the I.P.s, with the exception of the *Dis.*, experience a small improvement, that can be explained by a regularization derived from the discretisation of the head directions, thus eliminating some noise from the manual annotation since this process involves some errors due to low image resolution and small size of persons in the scene. In fact, the *Int.* profile is the one that presents the higher improvement, since is the one with less gaze variation, just corroborating the previous conclusion. However, the *Dis.* profile shows a performance drop since, by definition, it should present a high variability in gaze direction, thus

affecting its performance due to the imbalance of classes previously stated. Considering G.B.s, a small degradation is confirmed in all of them.

4.7.5.5 Mini-batch Approach

As stated in Section 4.6.6, the *fine mini-batch* approach uses the bags as samples and the final classification result of the whole individual trajectory (I.P.) or the entire group set (G.B.) is obtained considering the most predominant label along all the bags. An evidence score is taken as the ratio between the number of predominant labels and the total number of bags. The main theoretical advantage of the *fine mini-batch* over the *coarse mini-batch* approach is its dynamic nature, since if not prior knowledge about the starting and ending time of any I.P. or G.B. is known, the *coarse mini-batch* approach is useless, while the *fine mini-batch* maintains its formulation and can be used to detect the automatic switch between different I.P.s or G.B.s. Therefore in this section, we examine the robustness of the *fine mini-batch* level in order to suppress the *coarse mini-batch* level.

		E.I.			B.I.			U.I.			CHAT.			Avg.		
		P	R	A	P	R	A	P	R	A	P	R	A	P	R	A
MT-MG	Baseline	68.5	81.7	63.7	18.8	10.3	80.8	23.5	14.4	76.5	10.0	10.4	94.0	30.2	29.2	78.8
	Chamveha	91.7	89.4	85.1	**57.5**	55.0	92.2	40.5	48.3	88.2	13.3	21.1	95.2	50.8	53.5	90.2
	Our	**93.3**	**90.8**	**87.4**	48.4	**55.1**	**92.5**	**44.3**	**50.7**	**88.4**	**29.0**	**28.4**	**95.4**	**53.8**	**56.3**	**90.9**
AT-AG	Baseline	46.6	79.5	49.7	**21.1**	8.7	75.1	**30.9**	14.9	71.4	**16.3**	6.4	87.5	28.7	27.4	70.9
	Chamveha	**75.6**	80.5	**66.9**	15.1	**21.5**	**87.6**	24.8	16.8	**78.3**	**16.3**	**13.6**	**93.4**	**33.0**	**33.1**	**81.6**
	Our	73.4	**81.5**	66.4	12.9	13.9	86.3	28.0	**17.2**	78.1	11.8	6.8	91.6	31.5	29.9	80.6
		Dist.			Exp.			Dis.			Int.			Avg.		
MT-MG	Baseline	5.7	18.6	89.3	88.7	92.1	82.9	0.0	0.0	**99.7**	25.8	15.2	91.0	30.1	31.5	90.7
	Our	**28.0**	**22.5**	**89.7**	**90.7**	**95.1**	**87.2**	0.0	0.0	99.5	**68.4**	**39.1**	**95.2**	**46.8**	**39.2**	**92.9**
AT-AG	Baseline	3.1	**28.1**	**91.5**	**88.3**	91.3	**81.7**	0.0	0.0	**99.6**	25.9	**9.2**	**88.3**	29.3	**32.2**	**90.3**
	Our	**13.4**	9.2	86.9	80.9	**92.7**	76.8	**20.0**	**19.5**	99.4	**26.9**	7.2	86.3	**35.3**	**32.2**	87.4

Table 4.11: Classification results (%) for *fine mini-batch* approach.

Table 4.11 shows the classification results for I.P.s and G.B.s under both combinations of manual and automatic features extraction. For the G.B.s and assuming manual annotation, we confirm a clear advantage of both multiscale histogram descriptors, Chamveha and ours, over the baseline. *E.I.* and *CHAT.* behaviors are the most well-defined. It is expectable that for the *E.I.* behavior the performance difference between the baseline and remaining descriptors to be the smallest one. The overall low performance on *CHAT.* behavior can be explained by the small number of samples. In overall, our descriptor presents the best results over all the G.B.s and has a better recall rate that should be emphasised rather than the precision rate for surveillance systems.

Considering automatic features extraction, the baseline performance decreases less than both multiscale histogram descriptors, which suffer a high marked drop. However, their results are even better than the baseline with manually annotated data. This shows that the discretization of multiscale histogram can be affected and confused by tracking and gaze estimation errors. For its part, the small reduction undertaken by the baseline descriptor, which only includes the mean and

standard deviation of each feature, proves that our descriptor covers a good selection of discriminative features to describe individual interactions within a group, and that a global measurement that includes single occurrences could be representative enough to identify a collective behavior. Our descriptor is highly affected probably because the tracking and gaze estimation failures introduce noise that is captured by the multiscale histogram, which is cancelled out by the smoothing of the baseline. The Chamveha descriptor slightly superimposes to our descriptor, therefore we may conclude that since it has more features they likely complement each other in the presence of extraction failure. However, it also sustains the importance of our descriptor sampling strategy as an effective representation over time. In this scenario, the Chamveha descriptor presents a higher overall recall rate.

In terms of I.P.s analysis, our descriptor presents the best overall result for both manual and automatic feature extraction. However, we observe the same drastic reduction of our descriptor and just a small decay of the baseline while using automatic features. The *Int.* profile is the one with the worst degradation for our descriptor, probably because its dependence with the gaze feature and it is the one which presents the most structured movement aligned with the objects of interest of the scene, therefore perturbations on tracking and gaze estimation affect its performance. A curious factor is revealed, that the *Dis.* presents a significant improvement. Since this is the profile with the highest variation in gaze and position, the multiscale resolution of our descriptor is able to capture this behavior, while the baseline keeps its inefficiency in classifying this profile. Another reason behind this improvement, while using automatic features instead of manually annotated, is that the tracking and gaze estimation should emphasise the *Dis.*' profile characteristics, which are compacted at the bag level. However, we stress the fact that such profile has a very small number of samples.

		E.I.			B.I.			U.I.			CHAT.			Avg.			
		P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F_1
MT-MG	coarse	90.8	**93.9**	**87.9**	**52.7**	**56.8**	91.9	**56.6**	42.7	**89.6**	**35.6**	27.0	**95.6**	**58.9**	55.1	**91.2**	**58.0**
	fine	**93.3**	90.8	87.4	48.4	55.1	**92.5**	44.3	**50.7**	88.4	29.0	**28.4**	95.4	53.8	**56.3**	90.9	57.3
AT-AG	coarse	**79.2**	78.4	**67.6**	10.2	11.8	84.1	14.7	16.0	**80.9**	11.2	**8.0**	**93.6**	28.9	28.5	**81.6**	33.5
	fine	73.4	**81.5**	66.4	**12.9**	**13.9**	**86.3**	**28.0**	**17.2**	78.1	**11.8**	6.8	91.6	**31.5**	**29.9**	80.6	**33.6**
		Dist.			Exp.			Dis.			Int.			Avg.			
MT-MG	coarse	**28.9**	**44.1**	89.4	**95.3**	89.8	86.8	**11.6**	**10.0**	98.5	43.2	**54.2**	**95.7**	44.7	**49.5**	92.6	**55.6**
	fine	28.0	22.5	**89.7**	90.7	**95.1**	**87.2**	0.0	0.0	**99.5**	**68.4**	39.1	95.2	**46.8**	39.2	**92.9**	42.5
AT-AG	coarse	**13.6**	**23.2**	85.3	**92.3**	79.0	75.3	**24.6**	**42.5**	89.7	7.3	**21.7**	**87.4**	34.5	**41.6**	86.7	38.4
	fine	13.4	9.2	**86.9**	80.9	**92.7**	**76.8**	20.0	19.5	**99.4**	**26.9**	7.2	86.3	**35.3**	32.2	**87.4**	**42.6**

Table 4.12: Comparison results (%) between *coarse mini-batch* and *fine mini-batch* approaches, considering our descriptor.

Table 4.12 shows a comparison of performance between the *coarse mini-batch* and the *fine mini-batch* approaches. In terms of G.B.s both methods obtain similar results, with a slightly higher improvement margin for the *coarse mini-batch* approach while using manual annotation. A deeper analysis of each G.B. reveals an overall stable result for the *fine mini-batch* approach considering the automatic feature extraction. Therefore, we may conclude that the *fine mini-batch*

method is preferred since its performance is not significantly different than the *coarse mini-batch* approach, while having the advantage of being smoother and more stable.

Considering the I.P.s, an advantage of the *coarse mini-batch* approach over the *fine mini-batch* is verified when considering the manual data. However, for automatic feature extraction the *fine mini-batch* performs considerable better than the *coarse mini-batch*. In fact, the *fine mini-batch* approach keeps similar classification results with both manual and automatic data. As stated in Section 4.7.5.3, the I.P.s are more sensitive to noise and feature extraction failures than the G.B.s, therefore an obvious conclusion is that the *fine mini-batch* approach brings stability to the classification, since the collected information is more compact and consequently the descriptor becomes less error-prone.

	E.I.		B.I.		U.I.		CHAT.		Avg.	
	E.N.	E.P.	E.N.	E.P.	E.N.	E.P.	E.N.	E.P.	E.N.	E.P.
MT-MG	92.7	97.2	98.1	97.5	89.2	95.8	89.6	88.6	92.4	94.8
AT-AG	94.0	89.7	95.2	94.2	90.6	94.5	89.1	93.9	92.2	93.1
	Dist.		Exp.		Dis.		Int.		Avg.	
MT-MG	89.9	83.3	83.0	98.1	0.0	100.0	79.2	96.4	63.0	94.5
AT-AG	89.5	80.9	89.7	96.2	100.0	100.0	90.6	92.9	92.5	92.5

Table 4.13: Evidence scores (%) for false (E.N.) and true (E.P.) detections.

In order to obtain a measure of confidence for the decision of the *fine mini-batch* approach, we compute two scores: i) E.N., which is the *evidence of false detections* and indicates the level of confidence of false negative occurrences; ii) E.P., which is the *evidence of true detections* and gives the score of true positive occurrences. For an excellent decision process we expect a high E.P. and a low E.N. Inspecting Table 4.13, we state that our descriptor presents high values for both metrics, higher for E.P. than for E.N. This leads us to conclude that the *fine mini-batch* approach keeps a constant behavior along the entire trajectory, for the I.P., or the whole group, for the G.B. This regularity among the mini-batches emphasizes the discriminative power of this approach.

	E.I.	B.I.	U.I.	CHAT.	Avg.
MT-MG	90.6	48.1	46.8	27.3	53.2
AT-AG	66.7	14.6	30.3	11.8	30.9
	Dist.	Exp.	Dis.	Int.	Avg.
MT-MG	25.0	90.2	0.0	70.4	46.4
AT-AG	12.8	79.6	20.0	30.8	35.8

Table 4.14: Recognition rate (%) for extreme bags, initial and final, on I.P.s and G.B.s.

Since one the advantages of the *fine mini-batch* approach is its dynamic nature, which can be helpful for the automatic switch detection between I.P.s and G.B.s, we measure the recognition rate at the extreme bags, initial and final, of the I.P.s and G.B.s and reported them at Table 4.14. In general, we state that the recognition rate is directly related with the number of samples of I.P.s

or G.B.s, and there is an overall impairment of results for the automatic features extraction, with the exception of the *Dis.* profile.

4.7.6 *KRAV* Relevance Analysis and Classification

The proposed Kernel-based Relevance Analysis method, *KRAV*, formulated in Section 4.6.8, is used to obtain the vector $\boldsymbol{\varrho}$ holding the most relevant features from the original space $\boldsymbol{X}$. To this end, we adopt an iterative gradient descent algorithm for the optimization problem to find the projection matrix $\boldsymbol{A}$ shown in Eq. (4.11). Besides, to tune the Gaussian kernel bandwidth σ from Eq. (4.7), we use the information theoretic learning framework proposed by Álvarez-Meza *et al.* [16]. Given the high class imbalance exposed in Table 4.1, we set the resulting values from function $\delta_{ll'}$ (Eq. (4.8)) depending on the number of samples per class N_c as $\delta_{ll'}=1/N_c$ if $l=l'$, otherwise, $\delta_{ll'}=0$. As such, classes with few samples will be more relevant for the *KRAV* kernel-based learner, presented in Eq.(4.11). From matrix $\boldsymbol{X}$, we compute two new representation spaces $\boldsymbol{X}_s$ and $\boldsymbol{X}_e$, as explained in Section 4.6.8.

Considering the I.P. classification, we use $\boldsymbol{X}_s$ and $\boldsymbol{X}_e$ with a *k-nn* classifier with 10 folds and 10 iterations. The number of neighbors used for the classifier is heuristically found within the set $[1,3,5,7,9,11]$. We have selected the *k-nn* classifier within this branch of our framework, since it is the most suitable for some of the conducted tests, namely the continuous aggregation and the impact evaluation of individual features (see Section 4.7.6.2). To deal with the imbalance of the *Dis.* class during the cross-validation, we randomly select the 75% of the samples for training and the remaining 25% for testing in each iteration.

For the G.B. identification, the features are also encoded into the relational descriptor and the feature relevance analysis step is repeated in the same way as for I.P., resulting in the new feature spaces $\boldsymbol{X}_s$ and $\boldsymbol{X}_e$. Finally, the classification uses the same protocol as for I.P., but a cross-validation class balancing is not needed.

The evaluation of *KRAV* is carried out by means of the following three experiments: 1) **Relevance vectors analysis**, comparison and analysis of the relevance vectors $\boldsymbol{\varrho}$ obtained by *KRAV* and some state-of-the-art relevance analysis methods; 2) **Relevant features discovery**, calculation of a classification performance curve for each method while adding the respective features selected as relevant. This procedure is pertinent to find a new space $\boldsymbol{X}_S$. 3) **Relevant features embedding**, use of *KRAV* as a feature embedding tool to find new representation spaces from $\boldsymbol{X}_S$, in order to improve the classification performance and avoid the calculation of unnecessary features. To properly evaluate the impact of the above mentioned experiments into ***VISOBI***, we first use the low-level information, namely trajectories, gazes, and group formation/dispersion, from the annotation process, described in Section 4.4. This avoids the inclusion of noise derived from the automatic feature extraction algorithms. We should highlight that *KRAV* analysis is made at the bag level (see Section 4.6.6), therefore it is similar to the *fine mini-batch* approach in the BoF classification variant of ***VISOBI***. Next, the three experiments are explained in more detail, and the respective results analyzed and discussed.

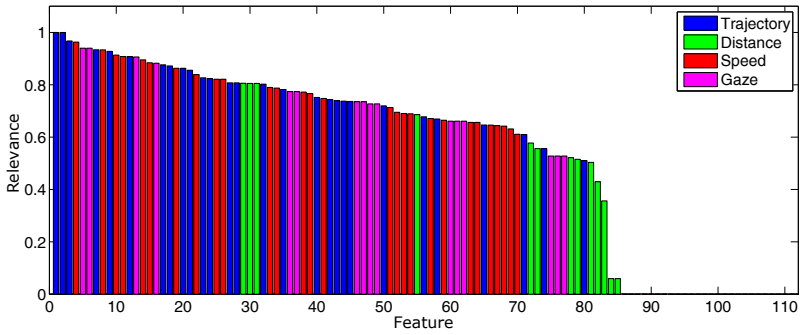
4.7.6.1 Relevance Vectors Analysis

We first use *KRAV* as a feature selection tool to support the understanding of the salient aspects of the input feature set, facilitating the interpretation of the employed spatio-temporal descriptors. To this end, we calculate the relevance vector $\boldsymbol{\varrho}$ ranking the original feature set of $\boldsymbol{X}$ as explained in Section 4.6.8.

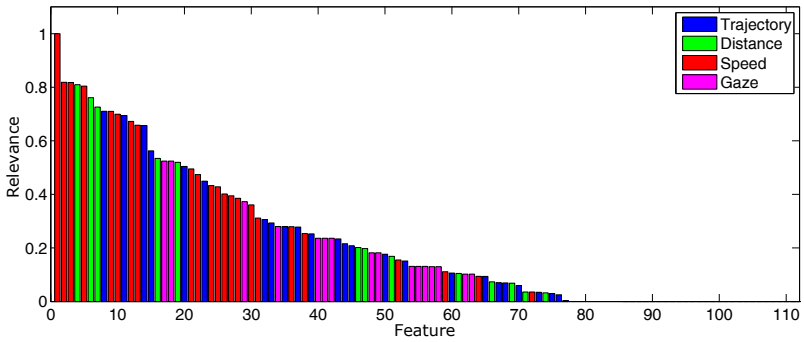
For the sake of comparison, the proposed *KRAV* is compared with two baseline feature relevance methods. The first one is a Variance-based Relevance Analysis (termed *VRA*), that ranks the input short-time features grounded on a variability criterion. Namely, *VRA* computes a relevance vector based on a linear transformation of the input representation space. Thus, *VRA* estimates the covariance among input features and the projection matrix maximizing the embedded space variability is fixed to computed such a linear transformation [77]. The percentage of retained variance parameter of *VRA* is set to 90.0%. The second baseline method called *Relief-f* calculates a relevance vector by looking to the closest of the same and different classes samples using a *k-nn* classifier [307]. The *Relief-f* parameter related to the number of nearest neighbors is set to one.

The obtained relevance vector $\boldsymbol{\varrho}$ using *VRA*, *Relief-f* and the proposed *KRAV* for the individual descriptor is shown in Fig. 4.16. Vectors $\boldsymbol{\varrho}$ are sorted in decreasing relevance order and normalized to the interval $[0, 1]$. As seen, the relevance vectors obtained by each method highlight a different set of features as relevant for the I.P. classification. For the *VRA* method, most of the features provide similar information as shown in Fig. 4.16(a). Besides, the bins related to the *distance* feature are not very relevant to classify the I.P.. The above can be explained due to *VRA* finds a linear combination of features, which maximizes the variability among data samples. However, in this case, the variability criterion might not be the most proper choice to enhance the separability among classes, since it does not include the supervised information of the labels. In contrast, the *Relief-f* and the proposed *KRAV* methods weight the features with a more discriminative order. This is given by the fact that these methods incorporate the labels information to rank the original input features. Regarding this, *Relief-f* finds more important the features related to the *speed* feature to improve the I.P. classification (see Fig. 4.16(b)). By the other hand, *KRAV* finds more relevant the features related to the *distance* as seen in Fig. 4.16(c). The difference between the obtained $\boldsymbol{\varrho}$ for these methods can be grounded by the fact that *KRAV* considers the high class imbalance for this classification problem, while the *Relief-f* gives the same importance to all the samples regardless the class membership.

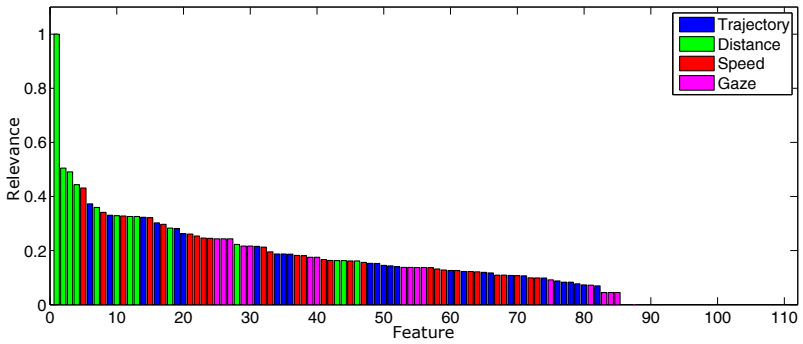
Further, for the G.B. relevance vectors shown in Fig. 4.17, it can be seen that *VRA* behaves similar as for the I.P., giving similar importance to most of the features. Moreover, the *profiles* feature information is not very relevant for the method. In contrast, as seen in Fig. 4.17(b) and Fig. 4.17(c), the *Relief-f* and *KRAV* methods identify as the most relevant features the *profiles* information and unlike for I.P., here these methods behave almost the same since there is not a big class imbalance. Some differences can be seen only after the 7th, where the *Relief-f* includes *speed variance* bins, while the *KRAV* recognize the *speed* feature as the most relevant.



(a) *VRA*



(b) *Relief-f*

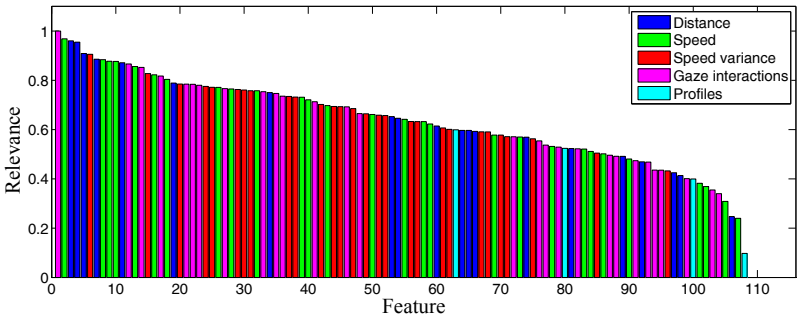


(c) *KRAV*

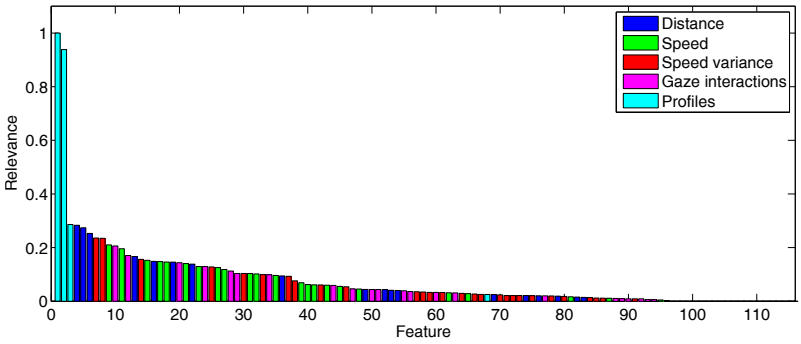
Figure 4.16: I.P. feature relevance analysis.

4.7.6.2 Relevant Features Discovery

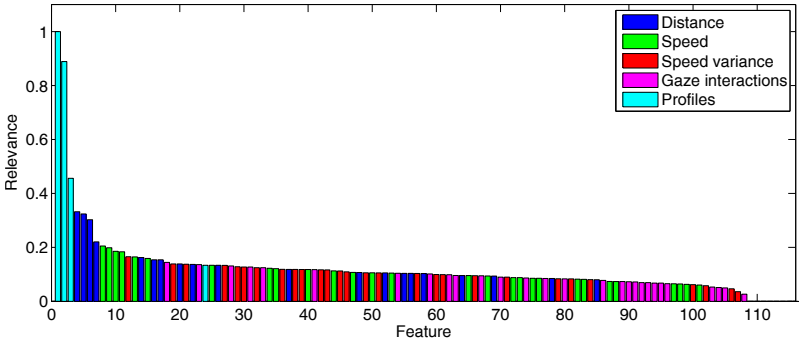
To inspect the impact of the features selected as relevant for the I.P. classification, we calculate the performance curve through the *k-nn* cross-validation scheme explained in Section 4.7.6. The classification performance for this experiment is assessed using the F_1 measure, which jointly consider the P and R measures. Fig. 4.18 shows the F_1 I.P. classification curve adding one by one the features ranked by the amplitude of $\boldsymbol{\varrho}$ for *VRA*, *Relief-f* and *KRAV*. The dashed lines indicate the selected subset of relevant features to conform $\boldsymbol{X}_S$ for each method, which corresponds to *VRA* $M_S = 78$, *Relief-f* $M_S = 41$ and *KRAV* $M_S = 23$. The threshold selection criteria to find M_S is set where the F_1 classification curve reaches the highest value.



(a) *VRA*



(b) *Relief-f*

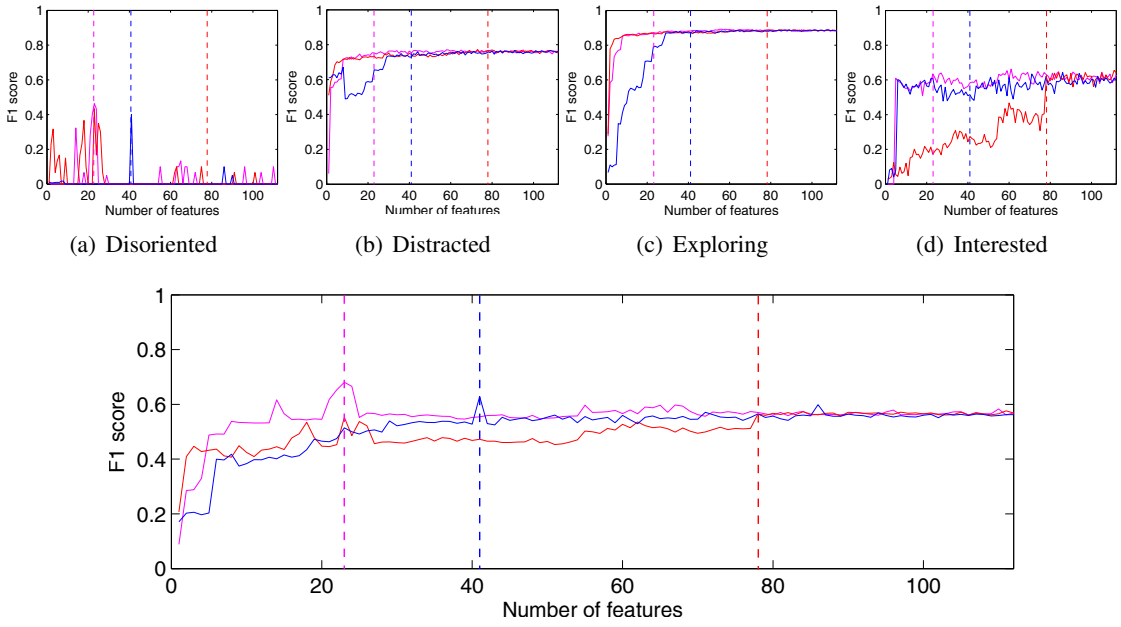


(c) *KRAV*

Figure 4.17: G.B. feature relevance analysis.

It can be seen from Fig. 4.18(e) that the proposed *KRAV* method obtains the highest classification performance of 68.1% with the lowest number of employed features. When analyzing individually the class performance, it can be seen that given the high imbalance, the *Dis.* class obtains the lowest classification performance, which only spikes when the selection of some particular features is obtained (see Fig. 4.18(a)). For the *Dist.* and *Exp.* classes, the highest classification performance is obtained, about 77.0% and 83.0%, respectively. Remarking that these values are obtained with the selection of a small set of features. Lastly, for the *Int.* class it can be seen that the *VRA* method has the worst performance, which can be explained by the low relevance given to the features related to the *distance* feature.

Same as for I.P, we show in Fig. 4.19 the G.B. F_1 performance curve adding one by one the features ranked by the amplitude of ϱ for *VRA*, *Relief-f* and *KRAV* as exposed in Fig. 4.17. In



(a) Disoriented (b) Distracted (c) Exploring (d) Interested

(e) Average

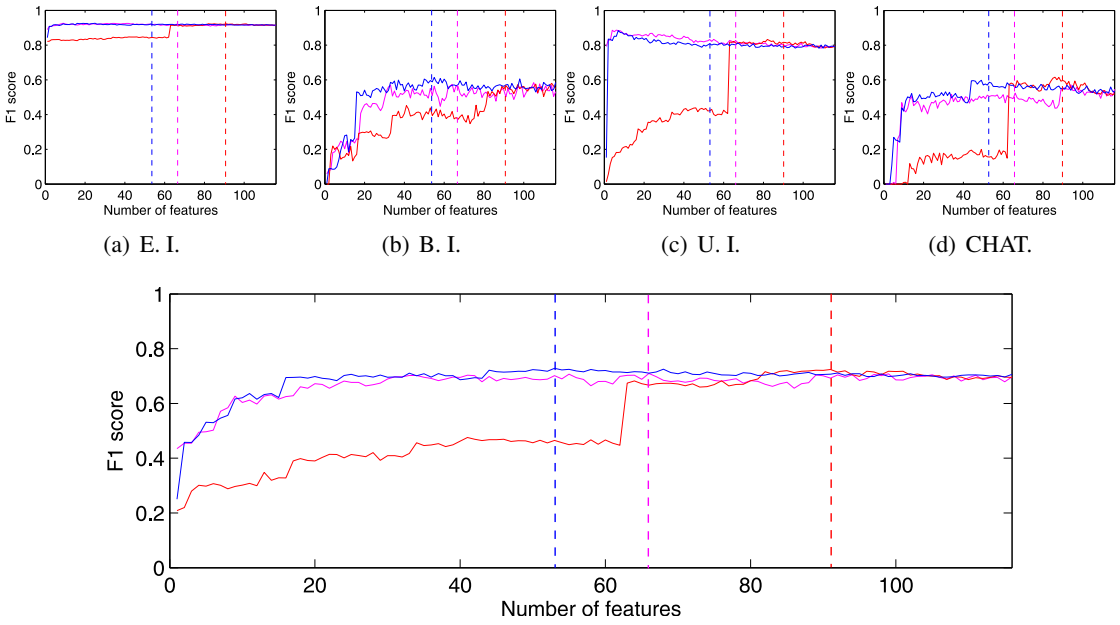
Figure 4.18: I.P. classification results while adding relevant features. —*Relief-f*, —*VRA* and —*KRAV*. The dashed lines indicate the selected M_S for each method.

this case, the selected subset of relevant features for each method corresponds to *VRA* $M_S = 91$, *Relief-f* $M_S = 56$ and *KRAV* $M_S = 66$.

The average classification results shown in Fig. 4.19(e) reveal that the performance curve for *Relief-f* and *KRAV* behave similarly, recognizing as the most relevant features the individual *profiles* and reaching a classification performance of 72.4% and 71.0%, respectively. By the other hand, the *VRA* method does not obtain a good performance until the 63rd feature is added, which corresponds to a *profile* bin (see Fig. 4.17(a)). This result demonstrates that the *profile* information is very relevant for the G.B. classification, as expected from the labels dependency stated in Section 4.4. When we individually analyze the classification per class, it can be seen that for *E.I.* and *U.I.* the obtained results are very high and are reached almost from the beginning of the plot for *Relief-f* and *KRAV*. This remarks that the *profile* features are discriminative enough to separate both classes. Differently, for the *B.I.* and *CHAT* classes the classification performance is not very high, and a larger number features is required to achieve a good F_1 measure. The latter can be related to the fact that these two classes have less samples the other two, thus, there is not as much information to properly learn the patterns to discriminate them.

4.7.6.3 Relevant Features Embedding

As a result of the above experiments, we obtain new feature spaces $\boldsymbol{X}_S$ using *KRAV* with the respective M_S for I.P. and G.B.. With the matrices $\boldsymbol{X}_S$, we perform two more tests using the proposed *KRAV* method, in order to measure the impact of the non-linear embedding of the proposed method into the classification performance. Furthermore, we want to analyze if it is possible to



(a) E. I. (b) B. I. (c) U. I. (d) CHAT.

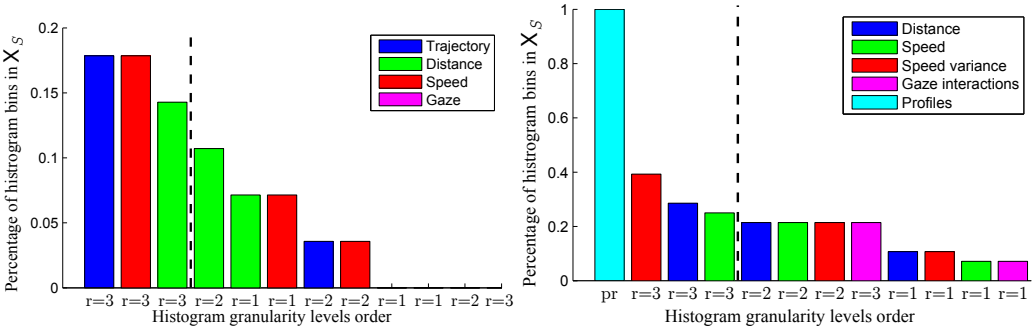
(e) Average

Figure 4.19: G.B. classification results while adding relevant features. —*Relief-f*, —*VRA* and —*KRAV*. The dashed lines indicate the selected M_S for each method.

avoid the calculation of some of the histogram features described in Section 4.7.6 maintaining a good classification. We performed the following evaluations for *KRAV*:

- *KRAV-S*: refers to *KRAV* as a feature selection, reporting the classification results for X_S exposed in Fig. 4.18(e) and Fig. 4.19(e);
- *KRAV-E*: we test *KRAV* as a feature embedding tool, which using the non-linear feature transformation, explained in Section 4.6.8, generates a new representation space $\boldsymbol{X}_E$ from $\boldsymbol{X}_S$ with the goal of improving the overall I.P. and G.B. discrimination performance;
- *KRAV-R*: we use *KRAV* to perform a non-linear embedding using a feature subset of $\boldsymbol{X}_S$. Namely, we create a new feature space $\boldsymbol{X}_R$ with the selected subset, which avoids the calculation of non-relevant histogram resolution levels for some features (see the I.P. and G.B. descriptor explanation in Section 4.6.5), while keeping a good discrimination performance. Fig. 4.20 shows the percentage of histogram bins per feature and resolution level in $\boldsymbol{X}_S$ for I.P. and G.B.. As seen, the obtained $\boldsymbol{X}_S$ for I.P. does not have any bin related to *gaze* information. This is caused by some errors and noise associated with the automatic gaze estimation method, which discretize the head orientation into 8 directions, and in our dataset we verify that such quantization is highly imbalance, since the constant flow of pedestrians is restricted to a diagonal path where the 1, 2, 5, 6 walking directions are very frequent, and the remaining almost negligible. Another problem is that the gaze estimation technique is highly dependent of the head orientation, but that does not guarantee to be an accurate estimation of the pedestrian view frustum. Moreover, the histograms with the most bins are

the ones of *trajectory*, *distance* and *speed* that correspond to $R = 3$ in our multi-scale descriptor. In this sense, we select these three histograms to conform $\boldsymbol{X}_R$ for I.P.. As for G.B., we conform $\boldsymbol{X}_R$ using the *profiles* and the same histogram bins that correspond to $R = 3$ for *distance*, *speed* and *speed variance* features. Again, *gaze* information is not found as relevant due to the previously discussed problem.



(a) I.P. feature subset (b) G.B. feature subset

Figure 4.20: Feature subsets obtained from $\boldsymbol{X}_S$ for I.P. and G.B.

Table 4.15 shows the average F_1 measure by classes for the I.P. and G.B. classification using the aforementioned training scenarios, plus the results obtained by *VRA* and *Relief-f* reported in Section 4.7.6.2. In general, we can observe that the average F_1 measure for G.B. is similar for all the considered methods, while for I.P. the results differ. This can be explained by the class imbalance problem, which is not considered by *VRA* neither *Relief-f*. Analyzing the methods individually, it shows that *VRA* obtains the lowest performance for I.P. (57.6%), because it does not take into account supervised information to rank the relevance of the input features. As explained before, the results for *Relief-f* improve in both classification (62.9%) and percentage of used features due to the inclusion of the labels information to find $\boldsymbol{\varrho}$. By its way, *KRAV-S* improves the I.P. results because it considers the class imbalance, obtaining 68.1%. Nonetheless, the results for G.B. slightly fall. Given the non-linear embedding, $KRAV - E$ obtains the best results for both I.P. and G.B., 74.8% and 76.1%, respectively. Lastly, the high $KRAV - R$ results demonstrate that the selection of the subsets illustrated in Fig. 4.20 are enough for a good classification, remarking that the percentage of used features is reduced from 21.0% to 12.0% for I.P. and from 57.0% to 25.0% for G.B..

4.7.7 Impact of Automatic Feature Extraction in *KRAV* Analysis

For this experiment, we analyze the impact of coupling together *GTIGC* and *KRAV* methods into ***VISOBI*** for social behavior identification. Additionally, we evaluate the influence of using the remaining feature extraction modules to automatically obtain the information related to the pedestrian detection and tracking, and gaze estimation. This experiment corresponds to the closest

	I.P.		G.B.	
	Avg F_1	# feat (%)	Avg F_1	# feat (%)
VRA	57.6	70.0	72.4	78.0
Relief-f	62.9	37.0	72.8	46.0
KRAV-S	68:1	21.0	71.0	57.0
KRAV-E	74.8	21.0	76.1	57.0
KRAV-R	73.9	12.0	72.3	25.0

Table 4.15: F_1 (%) results and percentage of relevant characteristics for the I.P. and G.B. classification using *VRA*, *Relief-f* and *KRAV* for feature selection and/or embedding.

real-world performance of the proposed ***VISOBI*** methodology with the *KRAV* classification variant. Following, the different tests for the proposed methodology are described and the attained results discussed.

- *VISOBI-E*: Based on *KRAV-E*, we test *KRAV* as a feature embedding tool to generate a new representation space $\boldsymbol{X}_E$ from $\boldsymbol{X}_S$. However, here we use the modules to automatically detect and track pedestrians, and gaze estimation, avoiding the use of the low-level ground-truth annotations. However, the group discovery ground-truth is still used.
- *VISOBI-R*: Based on *KRAV-R*, we test *KRAV* as a feature embedding tool using the same $\boldsymbol{X}_R$ feature space from *KRAV-R* to generate a new representation space. Nonetheless, as for *VISOBI-E*, we use the automatic tracking and gaze information, while using the group discovery ground-truth.
- *VISOBI-EG*: Same as *VISOBI-E*, but coupling the *GTIGC* method explained in Section 4.6.4.
- *VISOBI-RG*: Same as *VISOBI-R*, but coupling the *GTIGC* method explained in Section 4.6.4.

	I.P.		**G.B.**	
	Avg F_1	# feat (%)	Avg F_1	# feat (%)
VISOBI-E	50.8	21.0	68.6	57.0
VISOBI-R	49.0	12.0	62.1	25.0
VISOBI-EG	50.9	21.0	47.0	57.0
VISOBI-RG	49.0	12.0	51.4	25.0

Table 4.16: F_1 (%) results and percentage of relevant characteristics for the I.P. and G.B. classification using ***VISOBI*** for feature selection and/or embedding.

Table 4.16 shows the obtained I.P. and G.B. classification F_1 measures for the above four tests. It can be seen from the results of *VISOBI-E* and *VISOBI-R* that the inclusion of the algorithms for the automatic detection, tracking and gaze estimation seriously affect the classification performance, which drops about 24.0% for I.P. and 17.0% for G.B.. It is important to remark that the results for I.P. from *VISOBI-E* and *VISOBI-R* to *VISOBI-EG* and *VISOBI-RG* do not change since the group discovery does not affect the individual activities. As for G.B., the inclusion of the

automatic group discovery affects the classification results which drop about 21.0%. The slight advantage of *VISOBI-RG* over *VISOBI-EG* in G.B. classification is not expected. However, it can be explained by some noise included by the embedding and its posterior removal by the selected subset. The obtained results are still very relevant considering the high complexity of the dataset (see Section 4.4), which is given by the high confluence of pedestrians, supposing a big challenge for the pedestrian detection, tracking and group discovery methods. Furthermore, a last experiment was done to measure the individual impact of the *GTIGC* method in the G.B. classification. Therefore, considering the manually annotated data for the trajectories and gaze, and coupling *GTIGC*, we have obtained 58.5% for *KRAV-E* and 55.0% for *KRAV-R*. A difference of about 12% with embedding, clearly shows that the largest source of impairment of the system comes from the automatic tracking.

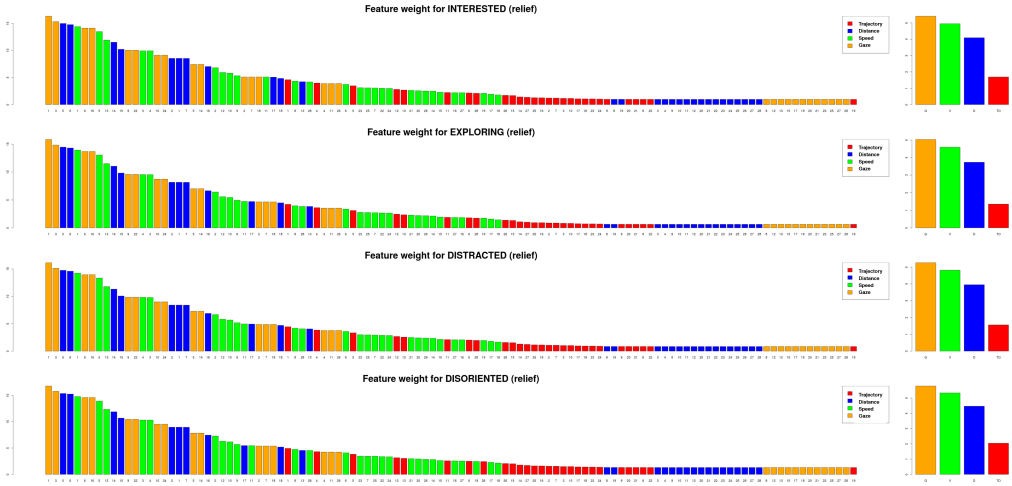
4.7.8 *BACK-F* Relevance Analysis

The proposed backward feature selection technique, *BACK-F*, explained in Section 4.6.7, permits to evaluate individual feature importance under the BoF classification framework. Inspection of Fig.4.21, shows that for I.P.s the features have well-defined and layered contributions, highlighting the relevant role of the gaze feature. On the other hand, G.B.s feature analysis shows a more balanced importance among features, proving their similar importance over the descriptor. This evaluation is conducted over the ground-truth information, in order to outcome the real contribution of each individual feature on the final classification.

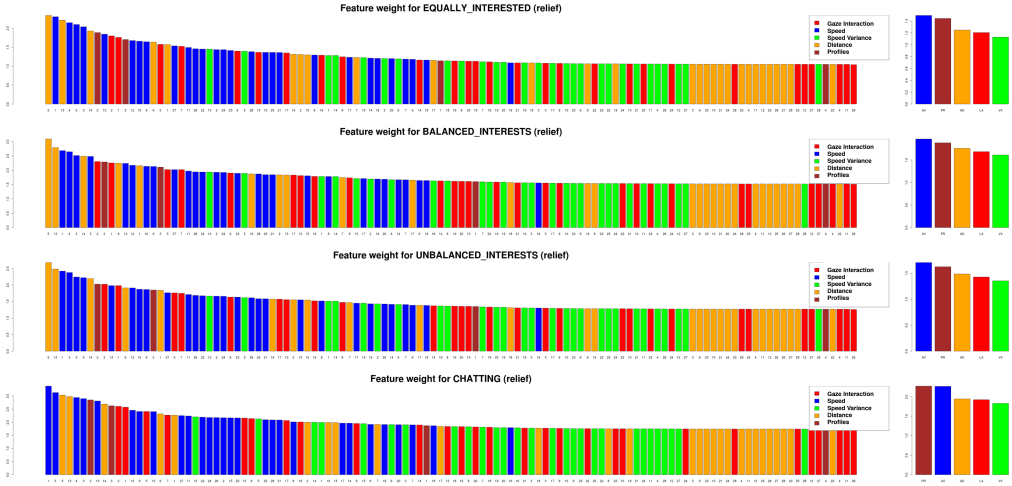
4.7.9 Sociological Meaning of Features

In this section, we inspect the sociological meaning of each individual feature within our descriptor and corroborate their importance for final classification as stated by both feature relevance analysis techniques described in Section 4.6.7 and Section 4.6.8.

Inspecting Table 4.17, and considering the G.B.s, we verify the importance of the individual profiles, P_p. By sociological definition [57], it makes sense since each profile enclose a well-defined behavior by itself, and the combination of all the individuals is what mainly defines the group behavior. However, higher values were expected since the definitions of our G.B.s concepts are highly dependent on individual profiles. This leads us to conclude that the manual annotation should be revised. Indeed, the annotation process is extremely hard, not only because of the difficulties regarding the low-level features such as gaze (affected by image resolution, camera viewpoint and perspective), but mainly because the decision to choose for the correct I.P. or G.B., which is a subjective process despite of the rules imposed (see Section 4.3). We also state that remaining features provide similar performance, confirming the feature importance analysis carried out in Section 4.7.8. Just the inclusion of the profiles, P_p, leads to a high performance rate, however, since they are also acquired from a similar classification process, noise and errors are introduced. In general, all the features contribute to the classification process and their mutual combination appears to be the most reliable option.



(a)



(b)

Figure 4.21: Feature importance analysis with *Relief-f* method for: (a) I.P.s; (b) G.B.s.

In terms of I.P.s, also a clear improvement is observed when all the features are aggregated. Indeed, this shows that they complement each other, which validate sociological theories that based their studies on several features to learn complex person behavior [57]. We verify important observations: i) for the *Dis.* profile all the contribution comes from the gaze feature, which makes sense since this is the profile with the highest gaze variability; ii) for the *Exp.* profile the predominant benefit comes from the speed, since it is the one that presents the lowest speed and this characteristic can help to discriminate it; iii) for the *Int.* profile the distance to nearest object of interest reveals the highest contribution, which in fact defines this profile. iv) for the *Dist.* profile all the features assume similar roles, since its behavior can vary depending on scene context.

As a reference of confidence, we conduct a final experiment to measure the features importance. We consider a two-fold stratified cross-validation scheme, where the class proportions are approximately equal on each fold, over 100 random iterations. In this way, we try to measure fea-

	E.I.			B.I.			U.I.			CHAT.			Avg.			
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F_1
$\tilde{v}_g$ + Var$[v_g]$	76.5	47.8	49.1	7.2	36.8	60.4	34.5	32.5	85.1	34.4	36.0	94.9	38.2	38.3	72.4	35.6
P_p	87.5	0.4	23.6	8.8	**91.8**	19.5	**78.1**	45.6	92.3	**73.1**	**100.0**	**98.5**	**61.9**	**59.5**	58.5	39.9
$\tilde{d}_g$	82.0	39.5	47.1	43.3	11.4	91.0	18.9	**57.4**	41.3	0.0	96.2	0.0	36.1	51.1	44.9	32.8
$laeo_g$	83.1	65.4	63.2	14.0	35.5	76.0	28.6	39.2	82.4	23.1	17.0	94.6	37.2	39.3	79.1	38.5
All	**90.8**	**93.9**	**87.9**	**52.7**	56.8	**91.9**	56.6	42.7	**89.6**	35.6	27.0	95.6	58.9	55.1	**91.2**	**58.0**
	Dist.			Exp.			Dis.			Int.			Avg.			
α_{si}	28.7	43.3	89.2	94.9	90.7	87.3	0.0	0.0	**98.9**	19.5	22.5	94.3	35.8	39.1	92.5	37.3
d_{io}	16.2	34.6	83.7	93.9	79.6	77.1	0.0	0.0	96.6	**43.2**	**75.3**	95.0	38.3	47.4	88.1	40.1
β_{gi}	14.4	29.8	84.0	92.0	76.4	72.9	10.5	**12.5**	98.7	6.0	19.7	86.7	30.7	34.6	85.6	37.5
v_i	**29.3**	42.3	89.4	**95.3**	**90.9**	**87.8**	0.0	0.0	98.4	24.7	27.3	94.6	37.3	40.1	**92.6**	38.1
All	28.9	**44.1**	**89.4**	**95.3**	89.8	86.8	**11.6**	10.0	98.5	**43.2**	54.2	**95.7**	**44.7**	**49.5**	**92.6**	**55.6**

Table 4.17: Classification results (%) of G.B.s and I.P.s considering combination of features within our descriptor, *fine mini-batch* approach and manual annotation data (see Section 4.6.5 for feature list).

	E.I.			B.I.			U.I.			CHAT.			Avg.			
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F_1
Stratified	71.4	69.0	84.5	**66.2**	**62.0**	81.7	54.1	**47.0**	75.5	**68.4**	**71.0**	82.7	**65.0**	**62.3**	81.1	**62.1**
Normal	**90.8**	**93.9**	**87.9**	52.7	56.8	**91.9**	**56.6**	42.7	**89.6**	35.6	27.0	**95.6**	58.9	55.1	**91.2**	58.0
	Dist.			Exp.			Dis.			Int.			Avg.			
Stratified	**37.2**	35.0	64.4	47.4	45.0	71.2	**50.0**	**25.0**	74.4	**57.8**	**70.0**	77.5	**48.1**	43.8	71.9	**62.6**
Normal	28.9	**44.1**	**89.4**	**95.3**	**89.8**	**86.8**	11.6	10.0	**98.5**	43.2	54.2	**95.7**	44.7	**49.5**	92.6	55.6

Table 4.18: Comparison results (%) for k-fold normal (keeping original classes proportions) and stratified cross-validation for our descriptor, *coarse mini-batch* approach and manual annotation.

tures importance of balanced data sampling. Table 4.18 shows that *CHAT* and *Dis.*, the G.B and I.P. with less samples respectively, largely improve their performance. We also notice, as expected, that the classes with more samples, namely *E.I.* and *Exp.*, decrease their classification results. The uniform distribution of samples among the training and testing sets improves, by a significant margin, the overall results for I.P.s and G.B.s, which states the importance of the selected features.

4.7.10 *KRAV* and *BACK-F* Discussion

Section 4.7.8 presents the results from the *BACK-F* method, which is supported by the sociological meaning analysis of the features conducted in Section 4.7.9. Briefly, that analysis states that for I.P.s the main importance comes from the *gaze* and the less significant is the *trajectory* feature. For G.B.s the importance of all the features is balanced, but we may highlight the *speed* and *profile* as the most significants.

However, the analysis pursued through Section 4.7.6 reveals a different behavior. For instance, in terms of I.P.s the major contribution comes from the *distance* feature, while for the G.B.s the *profile* is clearly the most important. In general, this analysis permits us to conclude that *gaze* is the less relevant feature for both I.P.s and G.B.s, and this represents the main inconsistence

among both analysis. Fig. 4.22 shows a summary of the importance of the individual bins of each feature's histogram, and the final importance of each feature, similar to the one shown in Fig. 4.21 in Section 4.7.8, for the *KRAV* method.

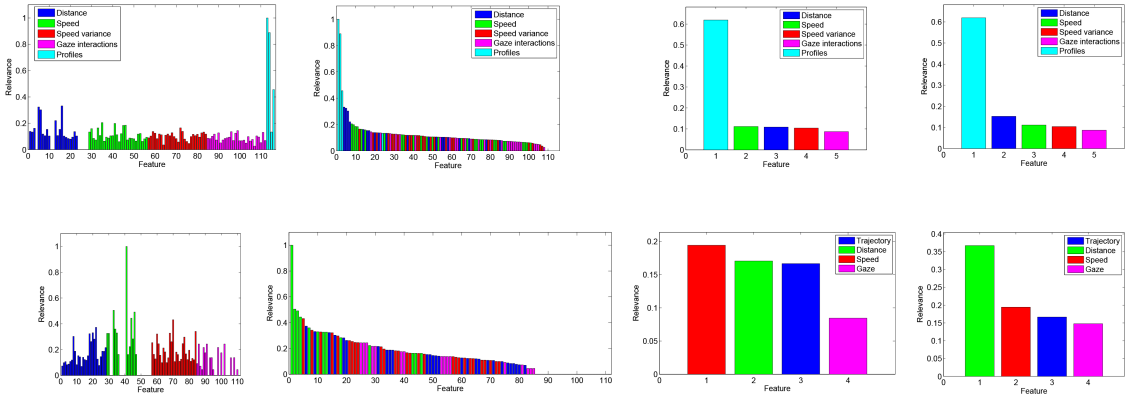


Figure 4.22: Summary of feature relevance analysis using *KRAV*, from left to right: relevance of the individual feature's bins; individual feature's bins orderly by increasing relevance; feature relevance (mean, considering the zero bins); feature relevance (mean, discarding the zero bins). Top row shows the analysis for G.B.s features, and the row below shows the results for the I.P.s.

In order to assess the discussion about the reason(s) about the disparity of results from both analysis, we may consider the *VRA* analysis conducted in Section 4.7.6.1, which states that *gaze* assumes an important role in the final classification for both I.P.s and G.B.s. This represents a similar behavior from the one obtained through *BACK-F*. The *VRA* method considers linear relations among the features and is grounded on a variability criterion, while the *BACK-F* uses the *Relief-f* method in the BoF representation. However, the *Relief-f* method is based on local neighbours from a *k-nn* classifier and incorporates the labels information to rank the original features. Since the *KRAV* method uses the CKA that consider non-linear transformations from kernel functions and also uses label information, similar to *Relief-f*, the only conclusion that we can draw is that the difference may come from the transformation caused by the BoF representation. Another reason, and more empirically, is that most of the individuals share a similar orientation, adding noise in the classification process, and therefore not being a discriminative feature. However, further analysis should be conducted in this direction. Our intuition is that *gaze* should assume an important role in the classification, however, comparing the results from Table 4.12, related to the BoF classification, and Table 4.15, related to *KRAV*, we verify a significant improvement, more than 20% considering the F_1 score, of *KRAV* over the BoF methodology.

4.8 Summary

In this Chapter, is addressed the characterization of individual profiles and collective behaviors within a social context in a surveillance scenario. For this purpose, new semantic concepts sustained on social-psychology principles are proposed and embedded into the annotation of a novel

video surveillance dataset for human activity recognition, validated by experts in the sociological field. This dataset represents a new perspective for the analysis of human activity within the computer vision community. Also, a complete framework, ***VISOBI***, is proposed that considers the social context inside the scene to assign high-level semantic labels to both I.P.s and G.B.s., and incorporates: a multi-resolution relational descriptor, automatic features extraction processes, including an improved group discovery algorithm, two feature relevance analysis techniques, and a novel two-fold classification approach based on mini-batches.

An extended evaluation process is presented. Mainly attention is given to the performance impact of the different automatic feature extraction steps into the classification results, and to the analysis of the sociological meaning of the individual features. It shows that automatic features extraction impair the classification results, especially from tracking despite the improvements added to the state-of-the-art Boosting tracker. The inclusion of pre-processing and post-processing techniques may help to improve each feature extraction process. The proposed group discovery algorithm, *GTICG*, unlike its predecessor, considers an improved merging mechanism for individual-group merging, includes a new property into the group coherence measure that brings more stability in the overall performance, and reduces computational effort without affecting the performance in video surveillance scenarios. However, the algorithm should be revised to increase its computational efficiency. As well, it should incorporate a mechanism to reduce the impact of tracking noise and error on its performance.

The classification approach can follow a classical BoF approach with different feature matching distances and pooling strategies, or can be conducted by the feature relevance analysis technique, *KRAV*, which uses two kernel functions to take advantage of the available joint information between the spatio-temporal features and the corresponding social behavior labels. *KRAV* handles in a better way class imbalance problems by selecting a set of relevant features and by the non-linear mapping of both individual and group classification. In general, the proposed mini-batch approach with different levels reveals promising performance and its dynamic behavior may be of great advantage for a real-time recognition framework of human activity in surveillance scenarios.

The global motion framework, ***VILOMA*** presented in Chapter 3, and the social behavioral analysis framework, ***VISOBI*** here presented, reveal solutions for two different problems in the same settings, i.e. surveillance scenarios. Their functionalities can be tie-together and their outputs can be shared in order to improve their performance on human activity-related tasks. Their complementarity comes from the integration of global motion representations with contextual features from local relationships among social entities in the scene. Chapter 7 reflects about their integration as part of a more complete solution within the *group* level of human activity analysis.

The next Chapter 5 changes the thesis's domain to the Multimedia settings. New context, different challenges to tackle. With the aim to study motion from different perspectives and aggregate context information from the background, multimedia video classification turns to be an excellent application scenario to study human actions as a *whole*. It presents a framework with individuals components that bring novel intuitions, based mainly in motion, to improve the final video classification rate.

Chapter 5

Multimedia Video Classification

> Vision is the art of seeing what is invisible to others.
>
> Jonathan Swift

Automatic video classification, indexing and retrieval require means to organize video data effectively. Video partitioning, camera motion compensation, detection of informative regions, spatio-temporal motion and appearance representations, and feature encoding are just few of the most important tasks in video classification [391]. In this problem, the categorization of the videos may be defined by human actions or by events, which, in the literature, implies an approach related to action recognition or event detection, respectively. Both approaches are closely linked since the detection of events involves the classification of actions and scenes, therefore, is normal that they share common feature extraction and encoding methods.

5.1 Introduction

Action recognition in videos has attracted a large attention in the computer vision community during the last years [3]. The recognition task involves the extraction of enough spatial and temporal information to distinguish the spatio-temporal patterns involving human movements in different activity categories. A related and more complex task is the detection of events in multimedia videos, which has obtained increasing interest in the area of multimedia in recent years. Indeed, Scherp and Mezaris [323] describe six aspects of events, namely *participation*, *interpretation*, *documentation*, *correlation*, *causality*, and *mereology*, which state their complexity by the inclusion of objects, subjectivity, correlation and causal relationships among events.

Event detection is accomplished in larger temporal videos than the ones used for action recognition. For instance, the most demanding event recognition task is the Multimedia Event Detection (MED) in TREC Video Retrieval Evaluation (TRECVID) [1]. In order to tackle multimedia event detection, the approaches used for action recognition can be applied directly, accompanied by the

[1] http://trecvid.nist.gov

inclusion of additional multimodal features such as audio and semantic concepts learned from textual event descriptions. In this way, action recognition systems may represent the visual analysis module of a wider event detection system.

Some of the big challenges involved in automatic video classification are the high intra and inter-category variations, and the uncontrolled capturing conditions. To tackle these problems, the visual analysis of this type of scenarios includes the representation of several entities statically, in each frame, and dynamically, throughout the temporal domain, within high camera viewpoint variance. Such entities may be known in advance, in case of the selected approach explicitly detects objects in the frame. If a non-detected-based approach is used, the detection of informative regions should be conducted; for instance to distinguish foreground from background. The scene could also be modelled in order to extract context knowledge. At the end, representations from different sources are encoded and combined in the best possible way to maximize the classification performance. This aggregation of multi-modal information is possible since the multimedia video content is extremely rich in terms of features' channels, and is what makes this scenario appropriate to the study of human actions at the ***whole*** level, defined in this thesis as the complete representation of the *frame*.

Under this settings, the previous frameworks proposed in this thesis are not conducive, due in particular to camera motion and viewpoint. Indeed, the ***VILOMA*** framework, presented in Chapter 3, assumes a static camera for the extraction of global motion patterns and the ***VISOBI*** framework, in Chapter 4, deals with a scenario where individuals are far away from camera, without close-ups, for the detection of individual and collective activity. Therefore, this Chapter presents our analysis of different motion and contextual representations for human action characterization in multimedia videos, denominated by *videos in the wild*.

Our proposal is based on an automatic visual detection framework, named **VIdeo-based MUltimedia Action Recognition** (***VIMUAR***), that aims to classify videos by human actions, while at the same time provides analysis of camera motion behavior and video-shot summarization information to improve the effectiveness of classification performance. From an application point of view, such framework may enable a faster browsing and permit a more efficient content indexing and access of large video collections. In this way, the proposed framework is evaluated on the most common action recognition datasets presented in the literature and the performance impact of its individual modules are measured, as well as their accuracy on related and supplementary tasks. The motivation behind the proposal is the construction of a complete and multimodal video classification system that may be able to identify events, detect and localize actions, and retrieve dynamic summaries for each video stored in a large multimedia video collection.

5.2 Overview

Video classification frameworks are normally composed by four stages: low-level feature extraction, feature encoding, classification and fusion of results. This pipeline can be abstracted into

two general steps: representation and learning methodology. The literature has produce considerable progress in both stages. The current state-of-the-art on several benchmark action recognition datasets is obtained by Wang *et al.* [364, 366] through a trajectory-based representation, together with an improved Fisher Vector encoding [284]. However, several efforts have been made on both parts to improve the final classification results.

Other common encoding techniques such as the Bag-of-Words (BoW) are widely used in large scale event detection systems. Such representations can be used at the frame level [242], *keyframe-based*, or at the video level [150], *video-based*. For the former, normally employed for still image features, temporal information is not included in the model and important keyframes can be missed during the detection of relevant keyframes. For instance, the Nikon MED 2010 system [216] is build under the assumption that a set of keyframes conveys enough information for event detection, reducing the problem to a classification of static images. Other systems, such as the INRIA-LIM-VocR and Axes [86] uses fixed-temporal keyframes to extract low-level static features in TRECVID MED 2014. The vulnerability of the latter, normally used for motion features, is the integration of noise from video length variation and unrelated features of small segments of the entire video, which are not representative of the video class. The IBM's MED system in TRECVID 2010 [242] applies a video-based approach downsampling the video to five frames per second. Duan *et al.* [87] exploit a video-based approach for the spacetime (ST) motion features from consumer videos. A different approach suggests the localization of the event segment of interest in order to suppress the drawbacks from the two aforementioned methods. Xu *et al.* [393] present a method to find the optimal frame alignment at different levels in the temporal domain to recognize events in TRECVID 2005. Yuan *et al.* [404] propose a spatio-temporal branch-and-bound search to localize the segment where an action might happen. However, such approaches are not reliable enough to be used in complex multimedia events videos.

Apart from the global representation of the video, small modules of the system may improve or damage the final classification results. They also may help to find new and effective ways to represent the video robustly and efficiently in terms of computational efforts. Following, we present an overview of some of the most relevant modules of a general video classification framework, introducing the motivations for our individual contributions.

5.2.1 Camera Motion Estimation

Camera motion is an important feature in video sequences since it not only determines transitions between different scenes, but it may also gives important cues about the position of the objects of interest. However, it also introduces a global motion component to the frame that can hamper processing tasks such as background subtraction, object detection, segmentation, among others.

Wang *et al.* [366] estimate the global background motion by computing the homography matrix between two consecutive frames. They generate points candidates through the complementary combination of features extraction and motion vectors from optical flow. The features correspondences are matched by the nearest neighbour rule, and the sampling of the flow vectors is based on the good-features-to-track criterion [332]. They use the RANSAC method [98] to estimate the

homography, and then rectify the images to suppress background camera motion and enhance the foreground moving parts. However, their biggest improvement is the use of a person detector to remove the inconsistent matches due to human movements when they dominate the scene; for instance in a close-up. In this way, they eliminate feature matches inside the detected bounding boxes to estimate the homography in a more reliable way. However, the detection of humans in unconstrained environments is extremely difficult. To overcome the problem of miss-detections, they perform forward and backward tracking of the bounding boxes. However, such solution is not accurate enough to compensate global camera motion.

Jain *et al.* [144] decompose motion into dominant, assumed to be the camera motion, and residual components, related to the independent scene motion parts. They use a 2D polynomial affine motion model [253] to estimate the dominant image motion. Then, they correct the flow vector by subtracting the computed affine flow vector from the optical flow vector, thus reducing flow vectors in the background and inflating foreground vectors. The assumption that global dominant motion is due to camera motion is not always correct, e.g. close-up of moving objects. However, their most important conclusion is that camera motion contains complementary information for the recognition of certain action categories at the descriptor level. They also verify that the compensated flow permits to extract fewer trajectories while keeping the same performance rate. Jian *et al.* [148] adopt an implicit way to deal with camera movement. They introduce two reference points to alleviate the effect of camera motion and model the relationship between objects and background. They cluster the trajectories to obtain the dominant motion and the global motion reference point, and use each trajectory as local motion reference point for motion characterization that incorporates trajectory descriptors and pairwise relationships. The representations are expressed by trajectory-based pairwise relative motion, which is intrinsically robust to camera motion. Their approach is highly dependent on the similarity metrics used by the clustering algorithm, and it is not able to model and obtain useful camera information. Heilbron *et al.* [125] introduce a new set of visual features that represent the context of the action. They distinguish background from foreground and model the context information by a combination of appearance background features and encoding of global camera motion. A simple camera motion model based on frame-to-frame fundamental matrices is used.

We argue that motion patterns that represent relationships among moving objects and static background should be incorporated into the recognition system, especially when the videos are captured under unconstrained environments. In such conditions, the camera motion is severe and may be a good discriminative factor for certain action/event categories. Most videos are recorded with an intention and normally the camera follows relevant motion and objects of interest depending on scene context, thus laying a correlation between the camera motion and the portrayed human action. Section 5.3.1 provides an overview of our proposal.

5.2.2 Motion Saliency

Video representation is normally achieved by the collection of space-time descriptors extracted at certain locations either sparsely or densely. Dense sampling offers a richer description of the

scene including contextual information at the cost of a high computational effort and noise aggregation. Therefore, it is pertinent to identify salient scene areas to filter out distracting regions, without affecting the discriminative power of the representation and obtaining a significant reduction of computational time in the subsequent processing steps. Although there are several saliency algorithms in the literature, the optimal predictor of visual saliency is based on data from human viewers, where *fixation density maps* convert raw fixation data, from eye tracking, into an empirical saliency map [292].

Vig *et al.* [359] take several experiments to investigate the impact of different salience-based (biological and non-biological models) masking of descriptors in the task of action recognition. Several important conclusions should be highlighted: i) peripheral masking presents poor results indicating that peripheral information may add noise to the descriptors; ii) simple central masking provide almost similar results than more complex techniques such analytical saliency in professional videos, which deflect the user attention to the center of the frame; iii) random uniform sampling discarding up to 60-70% of densely sampled descriptors does not impair baseline performance, which could suggest that random global subsampling at different spatio-temporal scales may include most regions of the video; iv) using only 20-40% of the descriptors masked by empirical saliency achieves the highest recognition rate, clearly stating the importance of biological models; v) the combination of descriptors from masked and unmasked regions are complementary, but performance decrease as the mask coverage increases. As a major conclusion, they state that a large amount of densely extracted descriptors is unnecessary and may even impair the detection of action recognition.

Considering that the human visual system uses visual attention to separate relevant from irrelevant regions on each task, our aim is to predict a motion-based saliency that can segment the image into *relevant* and *non-relevant* regions. In this case, the concept of *relevance* is obtained based on the combination of visual saliency, from color segmentation and motion similarity, and the assumption of intention behind camera movement. As mentioned in Section 5.2.1, our intuition is that in multimedia videos the camera normally follows the objects of interest, which can help to identify the action/event class. Refer to Section 5.3.2 for a detailed description.

5.2.3 Temporal Segmentation

A video usually contains several shots, and each one is composed of continuous frames that are captured in one camera action or content type. Each shot may represent meaningful discriminative parts for the classification task, or may introduce uncorrelated noise to the current video class. Apart from the common *keyframe-based* and *video-based* approaches, Sang *et al.* [286] present a *segment-based* methodology, where the videos are divided into uniform sampling segments for feature extraction and classification. In order to take into account the semantic boundaries between consecutive segments, they consider a dense sampling strategy with overlapped segments. In general, their results show that the *segment-based* approach outperforms the *video-based* approach, verifying a significant improvement of the overlapping sampling technique over the non-overlapping. As final remarks, this work shows the pertinence of segment localization for complex

video classification, however the segment-length varies depending on the dataset, and the overlapping sampling is not scalable since it requires a extremely high computational effort.

In order to segment the video dynamically, shot boundary detection algorithms may be used to identify the temporal boundaries between adjacent shots. There are several kinds of boundaries but they can generally be categorized into two types: *abrupt transition* (CuT) and *gradual transition* (GraT). CuT is usually easy to detect since the change on the boundary is great, and it can be associated either to a camera or content change. GraT is characterized by a smooth transition from one shot to another, which makes it more difficult to determine the starting and ending frames of the boundary. Depending on content and editing effects, it can be further divided into dissolve, wipe, and fade (out/in). The transition may also be confused by fast camera movement in a single shot, since the variation of content in both cases is smooth.

The usefulness of shot boundary detection techniques to improve event detection systems may not be clear and trivial. Sang *et al.* [286] apply a shot boundary technique [115], but they only rely on the automatic segmentation of scene cuts, discarding the results associated to the detection of scene transitions such as fade (in/out) or wide. They state a very low classification performance for their solution based on the shot boundary detection algorithm to extract segments, and conclude that the adopted technique is inaccurate on uncontrolled capturing videos. However, even in the presence of an accurate boundary detection algorithm, it is not intuitive how to combine multiple shots to improve the final classification. Indeed, the detected shots may have different lengths and may represent different contents that can be highly or lowly correlated with the video category. In this way, two questions arise: which is the optimal segment length, and how to detect and combine the representative segments to improve the detection of the video category.

A large active research in the detection of frame transitions in video sequences [157, 195, 405] has shown satisfactory results on the detection of cut and gradual transitions but without a fair comparison of results in the same dataset. For that purpose, TRECVID organized the shot boundary detection (SBD) task from 2001 to 2007. Several research groups participated on this task revealing interesting conclusions, targeting good practices to tackle the problem, and reaching competitive results, even with very difficult data from multimedia videos [334]. However, they were focus-oriented to the correct localization of the transition boundaries, while in our case we aim to investigate the detection of relevant shots for the video classification task. This is the main reason to propose a new method for video temporal segmentation.

Normally three steps are required for shot boundary detection: feature extraction and frame representation, temporal continuity measurement, and detection of location and type of shot boundaries. The first step is executed every frame and its representation can be pixel-based, histogram-based, descriptor-based or model-based. Cues such as color, edge, motion, or a combination can be encoded within the representation. The continuity step can be executed in a streaming approach, where a variable time window is used, or can be globally determined with the entire information of the video. The final step is normally computed immediately after the previous one using a predefined or adaptive threshold, a trained classifier, or a probabilistic model to segment the similarity curve.

Our proposal detects the shot boundaries caused by abrupt transitions, gradual transitions namely fade (in/out) and dissolve, and transitions caused by significant camera movement. More than just focusing on the accurate detection of the starting and ending frames of the transition boundaries, our solution looks for an alternative to summarize video content into meaningful shots with enough temporal duration, and intents to remove uncorrelated segments, which can be a possible source of noise. In this way, what is really important is the detection of the transitions, particularly the abrupt transitions, as they cause a change of content, and not to accurately detect their starting and ending frames. In fact, our proposal benefits from the earlier detection of the starting frame and the later detection of the ending frame of each transition. Those temporal shifts permit to assure that content presented during the transitions is completely discarded. Our intuition is that such content is not informative for any action or event category. Another important assumption in our solution is simplicity over complexity. Mas and Fernandez [213] have shown that a simple solution could obtain fair results even in demanding settings such as the TRECVID 2003 shot boundary detection task. Our solution integrates a streaming approach, to compute global similarities between two consecutive frames and compute candidate transitions along a sliding time-window technique, with a global approach to prune and rectify the detected transitions and identify valid correlated segments. Good effectiveness is achieved at significantly less than real-time processing. See Section 5.3.3 for more details.

5.2.4 Motion and Appearance Representations

Several works on action recognition have documented the importance of explicitly integrating motion characteristics into the dynamic video description. Indeed, motion is the most relevant cue to detect and identify actions of interest. One of the main issues in motion characterization is the separation of action motion from camera motion. However, this topic was already referred in Section 5.2.1.

As mentioned in Section 5.2, the current state-of-the-art work in video classification represents the motion by dense trajectories at different multi-scales [364,366]. Indeed, it is known that dense features perform better than sparse salient features for more complex videos [367]. Other major contribution of Wang *et al.* [364,366] is the introduction of descriptors computed in the space-time volume aligned with the trajectories. The MBH (Motion Boundary Histogram) descriptor is the best feature descriptors for these types of trajectories [366]. One of the reasons for this improvement is its robustness, to some extent, to the presence of camera motion, since it suppresses the constant motion by considering the flow gradient.

Following the representation of dynamic features, Histograms of Optical Flow [75] extract the velocity information of optical flow. A more complex motion descriptor, so-called Divergence-Curl-Shear (DCS) and that captures physical flow characteristic, is proposed by Jain *et al.* [144]. More precisely, it encodes the scalar first-order motion features divergence, curl and shear. Ali and Shah [14] exploit a set of eleven kinematic features computed from optical flow and use them to recognize spatio-temporal patterns denominated as kinematic modes. In terms of static appearance, one of the most common representations is the Histogram of Oriented Gradients (HOG) [74].

However, the combination of static and dynamic features permits to model the action into a more complete representation. Peng *et al.* [271], inspired by the co-occurrence of HOG (CoHOG) [375], propose a new set of spatial and temporal context descriptors that convey structural change of image (S-CoHOG, T-CoHOG), appearance changes along time (Co-HOF, T-CoHOF) and changes of spatial gradient orientations of flow (S-CoMBH, T-CoMBH). Sun *et al.* [347] use the velocity information of static and dynamic features to build a new descriptor (SDEV). Namely, the dynamic part provides information about the acceleration of intensity changes, and the static one captures the change of gradient with time. However, the works that explicitly concatenate static and dynamic features [148, 187, 201, 366], are the ones that achieve superior classification results.

The aforementioned methods only extract features on the dynamic trajectories belonging to the detected foreground, therefore they do not capture scene properties. Beyond local static and dynamic features extracted from foreground motion parts, the surrounding can be used to discriminate human actions and events. For instance, it may help to reduce the confusion between similar actions, and distinguish the context between events that present common motion patterns in unconstrained multimedia videos. Holistic models based on silhouettes [40] or motion [385] only perform well in controlled environments. Other approaches that include the whole video and measure the response to spatio-temporal filters [339], do not present competitive results when compared with the more recent approaches. The literature uses a very well-known sparse sampling Space-Time Interest Points (STIP) technique [186], and represent each point by the HOG and HOF of its local surrounding patch. SIFT (Scale Invariant Feature Transform) descriptor [204] is also often adopted to extract distinctive invariant features which make them robust to matching and, consequently, to accurately describe local patches. The GIST descriptor [257] is inspired on SIFT and creates a low dimensional representation of the scene as a whole from a set of perceptual dimensions. CSIFT (Color-SIFT) [356] presents a study about the analytical invariance properties of color within the SIFT technique. Reddy and Shah [302] state a significant improvement of CSIFT over SIFT, and a low performance of GIST on the UCF11 dataset. They also verify that the extraction of the spatio-temporal features is more effective over moving pixels than stationary pixels. Wu *et al.* [382] proposed the CENsus TRansform hISTogram (CENTRIST), which presents a high discriminative power to represent the structure of the scene. They state superior performance over the SIFT and GIST to recognize topological places and scene categories.

We argue that separating the background from the foreground produces a more reliable and robust context descriptor rather than describing the context holistically. The sampling frequency in which the context features are extracted can be important not only for the classification performance, but also for computational reasons. We propose an adaptive sampling technique based on information detected from saliency and camera motion type.

5.2.5 Features Encoding Strategies

The feature encoding stage turns sets of local features into fixed length vectors. The BoW model discards any structural spatial and temporal information among the features and for each descriptor applies a hard assignment to the closest visual word in the vocabulary obtained by clustering. A

word only captures information of a particular region and can be ambiguous for the classification task. Several extensions include soft-assignment techniques [94] or the use of spatial pyramids to take into account the image structural layout [188].

The Fisher kernel framework combines the benefit of generative and discriminative approaches and its superiority over the BoW model has been stated on image classification [284], and nowadays also preferred in video classifications tasks [342]. In this case, the features are modeled by their distribution as Mixture of Gaussians (GMM) forming a soft-codebook, and the data is represented by the gradient of its log-likelihood with respect to model parameters. The distance between the instances is performed with Fisher kernel. The dimension of the Fisher Vector obtained is $2KD$, where K is the number of clusters of the vocabulary, and D is the local feature dimension. The final vector representation is composed by two terms: one that considers the first order difference of feature points to clusters centers, and the other that integrates second order terms.

Peng *et al.* [272] present a comprehensive study about the global representation of BoW and the impact of fusion methods in several stages of its pipeline. They realise that each step of the BoW pipeline is crucial and recommend some good practices such as increasing the codebook size and a *hybrid fusion* at three levels: descriptor, representation and score. The same authors propose a Stacked Fisher Vector (SFV) [273] representation with multi-layer nested fisher vector encoding, inspired by the success of Deep Neural Network (DNN) for image representation and classification. Following the same deep structures to map low-level features to more abstract and semantic representations, Karpathy *et al.* [160] trained a Convolutional Neural Network (CNN) on a large collection of YouTube videos to take advantage of multiresolution spatio-temporal information. They state a significant performance improvement for action recognition on the UCF-101 dataset.

Under a multimodal classification problem, the combination of features, namely early and late fusion, is relevant to achieve a high recognition rate. There is no consensus in the literature on which one perform the best. According to Snoek *et al.* [338], late fusion tends to perform better, but it increases the learning effort. However, when early fusion shows best results, the improvements are more significant. Reddy and Shah [302] show that a late probabilistic fusion, which averaged the estimated probability of all the descriptors from their separately trained model, achieved the best performance when compared with early fusion. Hoai and Zisserman [128] learn a classifier score distribution over subsegments of the video, and verify that a final classifier based on this score distribution is more effective than using the maximum or average scores.

5.3 Action Recognition Framework

Our framework proposal, so-called ***VIMUAR***, is built upon the foundation of the state-of-the-art framework of Wang *et al.* [366]. Based on the intuitions and assumptions explained in the previous sections, as well as in the evaluation of individual system modules, our solution presents contributions in several stages of the framework with the goal to improve the final video classification

without damaging the entire computational system performance. Scalability is an important issue in this type of systems, therefore the re-use of modules and simplicity are good rules of thumb. Indeed, this type of systems are very complex and each module may introduce errors that are propagated to subsequent steps. This is also a motivation for the implementation of the individual modules in order to maintain a fair control over the system and inspect their individuals' impacts. Fig. 5.1 illustrates the complete pipeline of ***VIMUAR***.

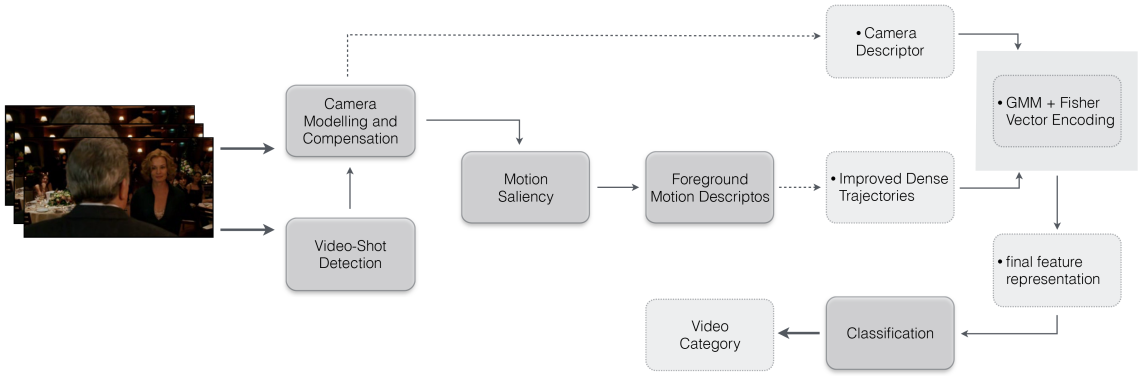


Figure 5.1: ***VIMUAR*** framework for video classification.

5.3.1 Camera Modeling and Compensation

Our aim is to model the camera motion for a two-fold purpose: i) compensate motion and, consequently, separate foreground from background; ii) estimate important physical camera parameters to describe the camera behavior. Figure 5.2 shows an overview of the main steps involved in this process, which are performed at each frame.

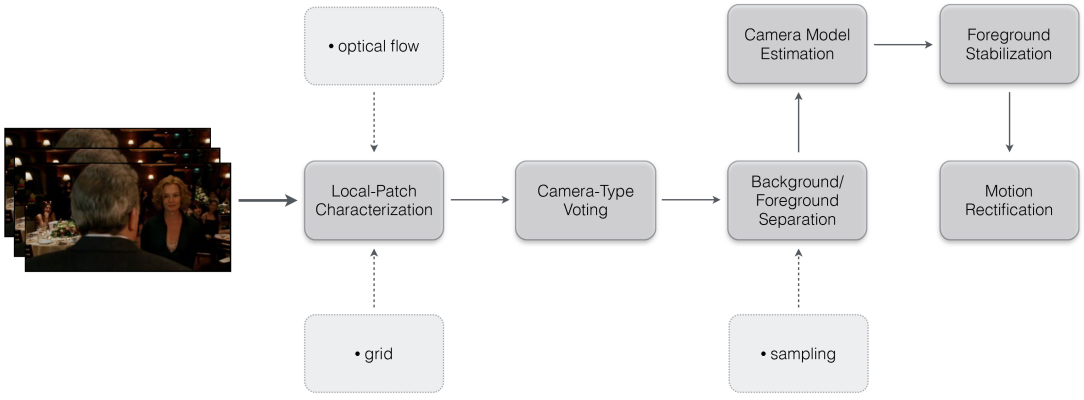


Figure 5.2: ***VIMUAR*** - camera modeling and compensation module.

The image is divided into a grid of $W \times H$ cells. At each frame t, the optical flow is computed and used to estimate the displacements of each cell $C(i, j) \in (W \times H)$ in x and y directions, $d_t^x(i, j)$ and $d_t^y(i, j)$ respectively. The displacements are treated as curves along a pre-defined temporal

window, τ. Since the camera motion may undergo large instantaneous displacement variation, the displacements are integrated through time to obtain cumulative displacements curves, $D_t(i,j) = \sum_{t=1}^{\tau} d_t(i,j)$, which cancel out small motion perturbations. In order to evidence long-term camera motions, the curves are smoothed by convolving with a Gaussian kernel as large as the size of the window. The curves are updated using a sliding window approach with overlapping to maintain continuity. A motion vector $M_t(i,j)$ at instant t is obtained from $D_t(i,j) - D_{t-1}(i,j)$.

As mentioned in Section 5.2.2, the camera normally follows relevant moving objects depending on the action/event class. When the human visual system focus a particular region of the scene, it creates a focus of expansion (FOE), which can be described by a high divergence motion field. Following the work of Poleg *et al.* [290], we estimate the FOE location by iteratively searching the location, considering only the central cells, that maximizes the radial projection response given by

$$\begin{aligned} \hat{R}_t(i,j) &= \begin{cases} 1, & R_t(i,j) \geq \cos(\theta) \\ 0, & \text{otherwise} \end{cases} \\ S_t &= \sum_{(i,j)} \hat{R}_t(i,j) \end{aligned} \tag{5.1}$$

$R_t(i,j)$, so-called radial projection vector, is the dot product between the normalized motion vector at cell (i,j), $\hat{M}_t(i,j)$, and the vector that connects the center of the cell with the FOE location.

Despite not being our goal to classify the camera motion type, i.e. static, translation, pan, tilt, zoom, an estimation of its global motion pattern can be a good rough approximation to separate foreground from background, especially if small local patches are used. The cells are used as local patches and the camera voting scheme presented by Ma and Wang [206] is adopted. Briefly, the idea is that each motion vector, $M_t(i,j)$, votes for a camera motion type, i.e. static, zoom, rotation and translation, and the last category is further subdivided into translational slices of $\pi/4 \in [0, 2\pi]$, considering the direction of $M_t(i,j)$ and a circle in the center of the frame. In our solution, we consider a primary motion type, static or translation, and a secondary type, rotation or zoom. The intuition is that videos in the wild normally present two camera motion types simultaneously, since they are captured by handheld devices, where translation and rotation are almost always present, and zoom and z-translation are confused. The vote is based in the following criteria:

i) if $|M_t(i,j)| \leq th_m$, th_m is a magnitude threshold, the vote is *static*;

ii) if $|M_t(i,j)| > th_m$, two votes are delivered: one for the *Translation*, accordingly with the direction of $M_t(i,j)$, and the other accordingly with $R_t(i,j)$: if $R_t(i,j) \in [\pi/2 - \varepsilon, \pi/2 + \varepsilon)$ is voted to *rotation*; if $R_t(i,j) \in [-\varepsilon, \varepsilon]$ or $R_t(i,j) \in [\pi - \varepsilon, \pi + \varepsilon]$ is voted to zoom ($\varepsilon = 0.6$).

The primary and secondary motion types with higher voting are selected, a voting confidence, $W_t(i,j)$, is obtained for the primary motion, which is normalized, and the weights for each $M_t(i,j)$ are updated as

$$w_t(i,j) = \begin{cases} w_{t-\lambda}(i,j) + \delta, & \in \mathcal{B} \\ \max(0, w_{t-\lambda}(i,j) - \delta), & \in \mathcal{F} \end{cases} \tag{5.2}$$

where $\delta = W_t(i,j) \times 0.5$. In this way, the background motion vectors ($\in \mathcal{B}$) will have large weights while the foreground ones ($\in \mathcal{F}$) will have small weights. This updating aids to stabilise the decision through time. If the voting confidence, $W_t(i,j)$, is not sufficiently large ($>th_{conf}$), the voting decision is discarded and motion camera types are kept. For the next frame, $M_t(i,j)$ is given by $D_{t+\lambda}(i,j) - D_{t-1}(i,j)$.

The same complementary sampling procedure of Wang *et al.* [366] is performed, i.e. SURF features matching plus flow vectors from good-features-to-track criterion. A coarse separation between foreground and background is made in a simple way: the samples that fall inside background cells are considered background, and vice versa. As any grid-approach, the size of the cell is a trade-off between the desired accuracy and the computational effort.

The selected background samples are used to compute the camera model. Considering a 2D affine model, defined at image point $p = (x,y)$ by

$$w_\Theta(p) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \tag{5.3}$$

the model parameters, $\Theta = (a_1, a_2, a_3, a_4, a_5, a_6)$, are estimated under a least-squares sense. The motion model parameter vector is transformed in terms of physical meaningful interpretation by motion sub-fields as divergence, curl, and hyperbolic such as $\Theta = (pan, tilt, div, rot, hyp_1, hyp_2)$ [42], where

$$\begin{aligned} hyp_1 &= 1/2(a_2 - a_6) \quad rot \quad = 1/2(a_5 - a_3) \\ hyp_2 &= 1/2(a_3 + a_5) \quad div \quad = 1/2(a_2 + a_6) \end{aligned} \tag{5.4}$$

Instead of a kalman filter to smooth the model parameters, a simple moving average is applied between frames. Afterwards, the foreground samples are stabilized by a global vector given by the translation T =(pan, tilt) and the rotation $\theta = \text{atan}(a_5/a_2)$. Finally, those foreground samples are used to compute the matrix to rectify the image, suppressing the background camera motion and enhancing the foreground moving objects. Instead of using the homography, the fundamental matrix is preferred since many scenes are not planar. The RANSAC algorithm [98] is used to remove the outliers, and then a refinement is applied using the standard 8-point algorithm [123].

5.3.2 Foreground-Background Saliency Maps

Our goal is to estimate a saliency map that represent relevant visual information to be used for further selection of foreground and background regions. We consider motion as the main feature

to define the concept of relevant, and adopt a contrast-based measure [276] to highlight the areas of interest. The steps are shown in Figure 5.3.

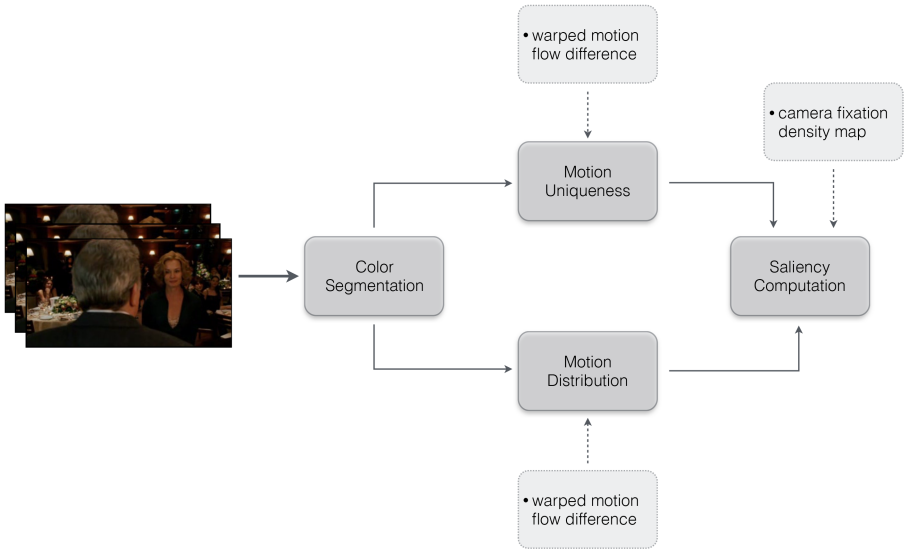


Figure 5.3: ***VIMUAR*** - motion saliency module.

An initial oversegmentation based on geodesic measure in CIELab space is performed. This permits to acquire small regions that preserve similar color structure and reduce the number of elements for subsequent processing steps. The uniqueness and distribution steps receive as input the warped motion flow difference. From camera motion modeling (see Section 5.3.1), the fundamental matrix is used to warp the optical flow and the difference between two consecutive warped flows is considered to be the warped motion flow difference.

The uniqueness factor is given by

$$U_i = \sum_{j=1}^{N} |m_i - m_j|^2 w(p_i, p_j) \tag{5.5}$$

where N is the number of segmented regions, m_i is the average motion in region i, and $w(p_i, p_j)$ is the local function that overemphasize the region boundaries in terms of spatial distance, and is modeled by a Gaussian such as $w(p_i, p_j) = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_p^2}(p_i - p_j)^2)$, where p_i is the average position of region i, σ_p controls the width of the operator, and Z_i is the normalization factor. The uniqueness defines the global singularity of each region in terms of motion.

The distribution factor is given by

$$D_i = \sum_{j=1}^{N} |p_j - \mu_i|^2 w(m_i, m_j) \tag{5.6}$$

where $\mu_i = \sum_{j=1}^{N} w(m_i, m_j) p_j$ is the weighted mean position of region j, and $w(m_i, m_j)$ measures the motion similarity between two regions i and j, and is also modeled by a Gaussian, $w(m_i, m_j) =$

$\frac{1}{Z_i}\exp(-\frac{1}{2\sigma_c^2}(m_i - m_j)^2)$. The distribution factor models the compactness of motion by spatial variance, indicating salient regions where the motion is coherent.

We have also verified that the D_i factor is highly discriminative, especially in our case where we are searching for compact motion components that may induce visual attention and express relevance for the action taking place in the video. Therefore, for the saliency assignment we also use the same formula as [276], namely

$$S_i = U_i \cdot \exp(-k \cdot D_i). \tag{5.7}$$

However, since we argue that the motion saliency is guided by the foveal view of the scene given by the camera movement, the scaling factor k is replaced by the average weight of each region i given by the camera fixation density map, $w_{\delta cam}$. This map is computed from the FOE location and the detected camera motion type (see Section 5.3.1), which is modeled by a two-dimensional Gaussian function fixed at FOE, such as $w_{\delta cam} = \exp\left(-\left(\frac{(x_i - FOE_x)^2}{2\sigma_x^2} + \frac{(y_i - FOE_y)^2}{2\sigma_y^2}\right)\right)$.

The final saliency motion mask is attained by an adaptive dual-threshold. It uses the mode of the S_i histogram as the mean value, and the thresholds are obtained by a deviation that increase at each iteration until a pre-defined percentage of the regions is assumed as foreground. This procedure is simple, low-resource consuming, since it is executed at region level and not at pixel level, and guarantees that the most salient regions are included in the foreground mask.

5.3.3 Video-shot Detection and Summarization

Our approach consists on four functional steps, as shown in Figure 5.4. The feature extraction is executed at each frame and represented by a combination of histograms. The inter-frame difference step is computed between consecutive pair-wise frames. The generation of candidate transitions follows a sliding window technique with overlapping, and finally the refinement detection of transitions and rectification of valid shots is computed at the end with the overall information.

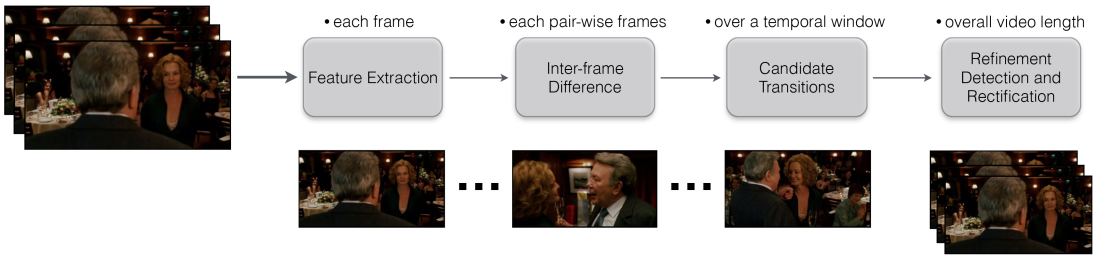


Figure 5.4: ***VIMUAR*** - shot-boundary algorithm module.

More in detail, for the feature extraction part the frame is represented by the combination of color, luminance and edge channels. HSV space is preferred for the color cue, where the V component, that correspond to the brightness, is replaced by the gray-scale luminance value, which is the mean of the RGB color space. This information is encoded into a three-channel histogram,

$h(c)$. The edge information is also encoded by a histogram, $h(e)$, that quantifies the contributions of HOG descriptors in several blocks, $b_x \times b_y$, over the entire image, and whose computation makes use of the fast integral image technique. The quantization applies a soft assignment between adjacent bins. Both representations follow a L_1 normalization.

Subsequently, the difference between two consecutive frames may be an indicator that a transition will occur at the present frame. Therefore, the intra-frame difference is calculated using the histogram intersection distance. Results for both color, $h(c)$, and edge, $h(e)$, histograms are equally weighted and the mean is taken as the global frame difference, $d_t = w_0|h_t(c) - h_{t-1}(c)| + w_1|h_t(e) - h_{t-1}(e)|$ (where $w_0 = w_1 = 0.5$). Considering a known threshold, this information may be enough to detect the high peaks associated to cut transitions, but not the gradual ones. Therefore, a sliding window approach with overlapping is adopted. During the temporal gap the intra-frame differences are stored and accumulated. Both curves, the one that keep the pair-wise differences and the other that represents the cumulative sum of the differences, $D_t = \sum_{t=1}^{\tau} d_t$, are smoothed by a Gaussian kernel. The continuity of the curves along the sliding window process is ensured by the overlapping factor. The cumulative curve provides a means of defining more clearly the gradual transitions, since their expected behavior can be almost modelled by a triangular shape (due to the applied linear video editing effects). However, this behavior depends on the motion presented during the transition, for instance large moving objects within the scene may cause the same shape behavior and, consequently, induce a false detection.

At the end of each sliding window iteration over the temporal domain, the search for candidate transitions is carried out. It is composed by three sub-steps: segmentation, pruning and linking. The segmentation uses the cumulative curve to coarsely separate segments with different slopes (positive or negative). Each segment keeps its slope sign, $\{s^+, s^-\}$, and the accumulated slope magnitude, $S_t = \sum_{t=i}^{j} d_t$. This step handles the continuity with the last segment stored in the previous sliding window iteration. The pruning step removes the segments that simultaneously present the number of frames and accumulated slope lower than some thresholds th_f and th_s, respectively. The last segment of each iteration does not undergo this step, since it will be processed in the next iteration. The linking step is the responsible to merge the continuous neighbour segments and detect the candidate transitions by a discontinuity criterion that considers four states based on the slope of the current segment, $\{S_{s^+}, S_{s^-}\}$, and on the type of the transition to detect, $\{S_c, S_g\}$ cut and gradual, respectively. The criterion is given by

$$\text{if } (S_{s^-} \text{ AND } S_c) = \begin{cases} |D_j - d_j| & > th_{max} \\ \sum_{t=j-1}^{i}(d_{t+1} - d_t) & > th_{max} \\ |D_j - D_{j-1}| & > th_{max} \end{cases} \quad \text{if } (S_{s^+} \text{ AND } S_c) = \begin{cases} |D_j - d_j| & < th_{min} \\ \sum_{t=i}^{j-1}(d_{i+1} - d_i) & > th_{max} \\ |D_j - D_{j-1}| & < th_{min} \end{cases}$$

$$\text{if } (S_{s^-} \text{ AND } S_g) = \begin{cases} \sum_{t=i}^{j-1}|D_{t+1} - D_t| & > th_{max} \\ |D_j - D_i| & > th_{max} \end{cases} \quad \text{if } (S_{s^+} \text{ AND } S_g) = \begin{cases} \sum_{t=i}^{j-1}|D_{t+1} - D_t| & > th_{max} \\ |D_j - D_i| & > th_{max} \end{cases}$$

where j is the last frame of the segment, i is the first one, th_{max} and th_{min} are the maximum and minimum difference thresholds, respectively. In case of $(S_{s-} \text{ AND } S_c)$, we expect a high and fast slope down, where the difference between D_j and d_j is high, and this corresponds to the starting frame of the transition. In case of $(S_{s+} \text{ AND } S_c)$, at the ending frame of the CuT-transition, is expected a stabilization which is reflected in a low difference between D_j and d_j. In case of $(S_{s-} \text{ AND } S_c)$ and $(S_{s+} \text{ AND } S_c)$ the behavior is similar, and the instantaneous is not taking into account due to its high variability. In these scenarios, it is expected a significant variation after some period of time, therefore we take into consideration the sum of D_j through the segment. In order to avoid false positive detections, the difference between the starting and ending frames of the segment is also measured. The threshold th_{min} is kept very low, and the th_{max} was tested empirically. We did not apply any of the adaptive threshold methods presented in the literature, since the obtained results are accurate enough for our intention.

The final step of the proposed shot-boundary technique aims to remove false positive and miss-detected transitions. Firstly, it verifies if there is any miss-detection by decreasing the threshold th_{max} by a percentage of the average of the maximum values at the inflection frame of the detected transitions. Then, it corrects the false positive detections by verifying the temporal continuity of the detected transitions.

The other goal of this module is to provide meaningful and correlated shots with the video category that their represent, as well as to remove uncorrelated shots that may impair the classification. Indeed, the detection of shots in video may induce an organizational temporal structure that most of the video recognition approaches do not use. This short summarization of the video can either be static or dynamic. A static summary implies the extraction of keyframes from each shot segment, and it must follow two main rules: i) it should be similar enough to the frames in the shot; ii) it should differ, to some extent, to the frames in other shots.

Figure 5.5: ***VIMUAR*** - shot-summarization step.

During the sliding window process, a uniform sampling is employed to store the frame and its histograms representations. At the end of the shot-boundary detection algorithm, those frames are collected into their enclosing shots. An *intra-shot* similarity matrix is created among the frames

of each shot. The frame with the highest average similarity is considered to be the representative frame of the shot. Then, an *inter-shot* similarity matrix is built among the shots using their representative frames. The shots are ranked by their descending order of similarity. It is important to mention that short-range shots are removed at the beginning of this process, therefore they are not processed. We set the minimum temporal duration of the shot equal to the maximum trajectory length used to represent foreground motion (see Section 5.3.4). This procedure is illustrated in Fig. 5.5.

5.3.4 Foreground Motion and Scene Features

The long-range trajectories presented in Chapter 3 are not suitable for this scenario. Their main drawback on this type of application is the computational effort involved. The flow advection and the 3D integration of the motion surface prevent their use in a large number of videos. Their effectiveness is reported as an useful technique to detect motion patterns and extract statistical measures of human behavior in several and different surveillance scenarios. They are extracted at uniform temporal mini-batches, but they lack temporal information in-between the starting and ending frames, which may impair the multimedia video classification since temporal information is omitted. The recording context is different which cause a different content context, i.e. in the surveillance settings the camera is fixed and the objects of interest are far away from it, while in the multimedia settings the camera is almost always in movement and the relevant objects are kept close to the camera frame. For these reasons, the streaklines and streamlines concepts are not useful in such a random and variable motion context.

In this way, the motion trajectories are computed as in [366]. After camera motion compensation (see Section 5.3.1), the trajectories are sampled and the descriptors computed from the warped image and flow, respectively. The trajectories just intent to represent the foreground motion parts, therefore, their dense sampling and continuity must lay within the foreground mask computed from our motion saliency model (see Section 5.3.2). This does not add any complexity to the system and removes a large number of points which are not in the detected foreground, thus reducing the computational effort. Also, for each trajectory the maximal magnitude of the motion vectors during its lifetime is computed to verify its overall displacement. If it is inferior to some pre-defined threshold, the trajectory is considered to be consistent with camera motion, and thus removed.

Considering the trajectory volume as $N \times N$ pixels and L frames, it is subdivided into a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$, where n_σ is the spatial subdivision and n_τ the temporal one. Appearance and motion based descriptors are aligned and computed within the trajectories volumes, and used to describe the foreground motion. Apart from the general descriptors used by Wang *et al.* [366], namely Traj. (trajectory information), HOG, HOF, MBH, we include the DCS descriptor defined by Jain *et al.* [144], since it brings important kinematic characteristics associated to flow, such as divergence, curl and shear. From their work, we may outline: i) in general all descriptors, Traj., HOG, HOF, MBH and DCS, benefit when warped flow (ω) is used to track the trajectories; ii) HOF computed from normal flow and warped flow are complementary and their

combination provides positive impact on classification; iii) MBH descriptor computation does not benefit from warped flow; iv) DCS is complementary to MBH, but by itself does not bring enough additional information; v) all the descriptors combined achieve, significantly, the best result.

The main conclusion is the advantage in using the warped flow to track the trajectories and compute the HOF and DCS descriptors. Since, we argue a more robust way to compensate camera motion and distinguish relevant foreground regions from non-relevant background regions, the inclusion of the DCS descriptor is pertinent, in order to evaluate our proposal and compare it with the reported by the authors [144]. Therefore, the foreground trajectories comprise the following descriptors: ω-Traj, ω-HOG, ω-HOF, ω-MBH, and ω-DCS. The size of a descriptor is given by $n_\sigma \times n_\sigma \times n_\tau \times n_b$, where n_b is the number of bins used by the descriptor. The size of the trajectory descriptor is $2 \times T$, since it includes the x and y displacements along the trajectory.

Scene context may provide additional information. In this work, as context we use the camera movement and explore the impact of our camera modelling module (see Section 5.3.1) into the final classification. The proposed camera motion descriptor, denominated by D_{cam} for simplicity, is a combination of the global physical motion captured by the motion model parameter vector Θ, plus statistical information collected by the local motion detectors, i.e. the cells. In this way, the descriptor also considers the radial projection response S_t and the average, μ, and standard deviation, σ, of the motion vectors of the peripheral cells, since they are the ones that may reflect the most predominant motion in terms of camera movement and as stated by Vig *et al.* [359], peripheral image regions are a complementary source for descriptor encoding. The final descriptor vector, D_{cam}, has a dimension of 7+2P, 6 (Θ) + 1 (S_t) + 2P (# of peripheral cells).

5.3.5 Features Encoding

Both BoW and Fisher Vector representations dismiss spatial and temporal information of the features. To overcome this drawback, we adopt the technique proposed by Sun *et al.* [342] that build a spatial-temporal pyramid structure. At the pyramid level i, the video is divided into 2^i slices along temporal domain and $2^i \times 2^i$ spatial blocks for each frame. Considering P_s spatial pyramid levels and P_t temporal pyramid levels, the total number of sub-volumes is given by $V_t = (4^{P_s} - 1)(2^{P_t} - 1)/3$. The encoding is performed for local features in each sub-volume, and the final representation is obtained by the concatenation of all vectors. Considering the best practices on image classification [284], each dimension of the fisher vector follows a power normalization, flattening the contribution of each dimension.

5.4 Validation

We use the state-of-the-art work of Wang *et al.* [364, 366] as the baseline comparison among the several steps of our proposal to measure their individual impact on the final classification. When necessary, other state-of-the-art works, including new deep-learning-based approaches, are included to compare the final performance.

5.4.1 Datasets and Evaluation Protocols

Hollywood2 benchmark [187] contains almost 6 hours of video data collected from 69 different Hollywood movies and split into 823 training and 884 test video clips. The set contains 12 actions and video clips may have multiple labels. This datasets captures well many challenges in action recognition, such as high intra-class variability, high degree of clutter and camera motion, and ambiguity in clip labelling. We use 50% of the dataset, maintaining the distributions of class instances. The performance is measured by mean average precision (mAP) over all classes, as in [212].

Olympic Sports [246] represent sports videos collected from YouTube and annotated using Amazon Mechanical Turk. There are 16 sports actions, represented by a total of 783 video sequences. The used train-test set is the one provided by the authors [246], i.e. 649 sequences for training and 134 sequences for testing. We use 50% of the dataset, maintaining the distributions of class instances. The performance is reported by mAP over all classes.

HMDB51 [177] is a relatively new dataset that aggregates 51 action categories from a variety of sources ranging from digitized movies to YouTube videos, in a total of 6,766 video sequences. We follow the original protocol using three train-test splits for the original videos. For every class and split, there are 70 videos for training and 30 videos for testing. We use 50% of the dataset, maintaining the distributions of class instances. The performance is reported by average accuracy (mACC) over the three available splits.

5.4.2 Experimental Setup

Following the procedure of [366], two subtle pre-processing steps are performed to the datasets: i) videos are normalized to have the height of 360 pixels; ii) frames are extracted at 25 fps. These modifications are added to standardize the feature extraction procedure across videos and datasets. They do not significantly alter the performance of the video recognition system, as verified in our experiments.

The parameters of the trajectories are set to default, i.e. N=32, n_σ=2, n_τ=3, L=15. For the descriptors we have n_b=8 for HOG, MBH, and DCS, and n_b=9 for HOF (see Section 5.3.4 for details). To encode features, we use Fisher Vector, as mentioned in Section 5.3.5. We first reduce the descriptor dimensionality by a factor of two using Principal Component Analysis (PCA). We set the number of Gaussians to K= 256 and randomly sample a subset of nearly 50% of features from the training set to learn the GMM. There is one GMM for each feature type. Considering the spatial and temporal pyramid levels described in Section 5.3.5, each video is represented by $2DKV_t$ dimensional Fisher Vector for each descriptor type, where D is the descriptor dimension after performing PCA, and V_t the total number of sub-volumes (for our case, we set P_s=1 and P_t=1, therefore V_t=1 for comparison purpose with the literature, since we mainly want to evaluate the contributions from motion compensation, foreground-background distinction and temporal segmentation).

The camera descriptor, D_{cam} is computed at uniform sampling of k frames, and at every frame that a camera motion type is detected. The D_{cam} descriptors does not follow any dimensionality reduction. It is encoded into a normal Fisher Vector for each sampled frame, normalized using power and L_2 normalization. Features from these sampled frames are averaged at the end to be part of the video level representation. Finally, to represent the video, we concatenate the normalized Fisher Vectors from the different descriptors types. In all the experiments, we use a linear SVM and fixed C=100.

5.4.3 Motion Compensation Impact

In this Section, we investigate the effectiveness of our technique in compensating the global motion background caused by camera movement (see Section 5.3.1). The most useful evaluation for our purpose, is to measure its impact in the final classification. Indeed since this framework module computes the rectified matrix to warp the optical flow, so-called warped flow (ω), it influences the trajectories tracking and the computation of the descriptors. As baseline, just our camera motion compensation and motion-saliency modules are replaced in the work of Wang *et al.* [366], which is referred here as *IDT* and used in the further experimental setups for sake of brevity.

Table 5.4.4 presents the comparison between the baseline and *IDT* with (WithHD) and without (WithoutHD) automatic person detector to remove the inconsistent matches. For *IDT-withHD* two results are presented: the one reported in the paper for the complete datasets, and the other computed for us in the same subset of the datasets used for our experiments. At a first glance we see a difference of results for the *IDT* approach between the reported ones and the obtained by us in the subset. The major difference is in the *HMDB51* dataset, probably because it is the most complex one in terms of content, and the one that presents more classes. Assuming a direct comparison of our baseline with *IDT* in the subset, we state a better overall performance of the baseline. It just presents a worse result in the *HMDB51* dataset, where recognition performance drops slightly compared with the *IDT*. What is remarkable, is its behavior in the *Hollywood2* dataset. No recent work has achieved, so far, results better than the one that we report here. From visual inspection of some scenes of the *Hollywood2* dataset, we state that most of the scenes are close-up of actors with the actions well-defined but with a clutter background. If so, our camera motion compensation approach improves the computation of the correct camera model in the presence of objects/person that occupy most of the frame. In the same way, the proposed motion saliency map is most effective and reliable than object-based detectors methods used to segment the foreground from background, as in [366]. These results provide evidence about the importance of camera motion compensation and the appropriate selection of regions relevant to the action being performed. A further experiment should be conducted to assess the contribution of each component in the final classification.

Method	Hollywood2	Olympic Sports	HMDB51
IDT-withoutHD (from [366])	63.0	90.2	55.9
IDT-withHD (from [366])	64.3	91.1	57.2
IDT-withHD (subset)	62.0	88.8	51.0
Baseline	71.1	89.9	50.4

Table 5.1: Comparison results (%) of our baseline, with camera motion compensation and motion-saliency modules included, to the *IDT* approach, with and without automatic person detector (mAP for *Hollywood2* and *Olympic Sports*, and mACC for *HMDB51*).

5.4.4 Motion Saliency Evaluation

The purpose of this Section is to clarify the effect of the motion-saliency module by constraining the sampling process of the trajectories, and, consequently, of the scene features. In this way, the frame is divided into background and foreground regions. For sampling the scene features, the background regions are used, and for sampling the trajectories the foreground regions are used instead.

In this experiment we just take into consideration the number of trajectories extracted, since the *IDT* does not extract scene features. Table 5.2 shows the comparison with *IDT-withHD* in the selected subsets. The total number of trajectories decreases significantly for sampling points just in the background areas detected by our motion-saliency module. Combining these results with the previous ones, from Table , we verify an overall improvement of classification, while reducing the number of trajectories. For sure, this also reduces the computational cost, enabling the computational resources for other tasks, such as the extraction of more demanding descriptors that may use spatial and temporal co-occurrences, as Peng *et al.* [271] have proposed. These results permit us to corroborate our intuition that motion is a very important cue for detecting relevant regions, and that is guided by the foveal view of the scene given by the camera movement.

Method	Hollywood2	Olympic Sports	HMDB51
IDT-withHD	169909273	146573244	222346819
Baseline	153267281	93470807	168316541
Diff	10.0%	63.8%	24.3%

Table 5.2: Comparison of the # of trajectories extracted from our baseline, with camera motion compensation and motion-saliency modules included, to the *IDT* approach, with automatic person detector. All the results consider the subsets of the datasets used in our experiments. The final row shows the reduction of # of trajectories in terms of percentage.

5.4.5 Video-shot Boundary Detection

In order to evaluate the accuracy of the proposed shot-detection algorithm, we measure its performance on three datasets:

i) *Otawa-CuTseg* [1], a dataset created for the detection of CuT transitions, with 10 videos of very different contents such as high-quality TV-shows, commercial sequences, news sequence, cartoon clip, low-quality TV video, and movies, among others. We compare our results with the author's results [378];

ii) *Hollywood2*, a very large dataset of different Hollywood movies with manual annotation of the shots presented in the 1707 video clips. We provide our results for the entire dataset;

iii) *TRECVID SBD 2007*, a very challenging dataset: 17 videos, in a total of 637805 frames, 2463 CuT and GraT transitions. We compare our results with some research groups that participated on this challenge.

The results are presented under three metrics: Precision (P), Recall (R) and F_1 measure. As stated in Section 5.3.3, our algorithm was desgined to detect earlier the beginning of the transition and later the ending of the same. This margin of error intents to remove all the frames within the transitions and boundaries, without discarding meaningful information of the detected shot. For this reason, the aforementioned metrics take into consideration a gap of 5 frames before and after the starting and ending frames, respectively. Therefore a transition is considered as correctly detected if

$$(F_{gt}^{l} - 5) > F_{d}^{l} \leq F_{gt}^{l} \quad \text{and} \quad F_{gt}^{u} \geq F_{d}^{u} < (F_{gt}^{u} + 5) \tag{5.8}$$

where F_d^l and F_{gt}^l are the detected and ground-truth starting frames, respectively, F_d^u and F_{gt}^l the detected and ground-truth ending frames, respectively. The *Frame Precision* and *Frame Recall* measures reported in the TRECVID SDB 2007 task [1] are not taken into consideration in our evaluation.

Table 5.3 shows that our method outperforms, by 1.6% in F_1 score, the best algorithm reported so far in the dataset *Otawa-CuTseg*. This improvement over the state-of-the-art states the robustness of our algorithm for the detection of CuT-transitions. However, the R measure is slightly smaller than the state-of-the-art. Inspecting the results for the individual videos, we notice that almost all the miss detections come from video J, which is a trailer of a film that presents many computer generated features, therefore many close proximity CuTs. This result makes sense, since the way our algorithms works, it is obvious that it will miss close transitions. However, this is not a drawback, it is a feature of the algorithm, since by detecting earlier and later the beginning and ending frames of the transitions implicitly it filters out the consecutive transitions. In this way, the shots in-between such transitions are not considered as valid ones, since they do not have enough meaningful information to be used for the classification stage.

[1] http://www.site.uottawa.ca/~laganier/videoseg/

[1] http://www-nlpir.nist.gov/projects/tvpubs/tv7.slides/tv7.sb.slides.pdf

Method	P	R	F_1
Pixel-based Loc.	77.4	80.0	78.3
MOCA [285]	97.1	70.1	79.4
Whitehead *et al.* [378]	87.4	**96.1**	90.8
Our	**89.6**	95.4	**92.4**

Table 5.3: Results (%) in the *Otawa-CuTseg* dataset.

We did not found any work that report the results of shot-detection for the *Hollywood2* dataset. However, since it is one of the datasets used to measure the classification performance, it is pertinent its evaluation for the shot-detection task. The results are high enough to trust in the generated shots for the classification stage. Comparing these results with the obtained in the *Otawa-CuTseg* dataset, we notice that there is a significant increase in terms of miss-detections and a slightly decrease of false positives. The difference between both datasets in terms of the number of videos is very large, from 10 to 1707, which is important to conclude that the algorithm keeps its performance in terms of false positive. Our intuition is that the increase in terms of miss detection comes from its behavior towards close transitions, which are common in this dataset since it is composed by Hollywood movies.

Method	P	R	F_1
Our	90.4	83.2	86.6

Table 5.4: Results (%) in the *Hollywood2* dataset.

Finally, Table 5.5 shows the results for the *TRECVID SBD 2007* dataset. All the results, except the ones from *Bradford Univ.* and *Sheffield Univ.*, are reported as the mean of all the runs they submitted. The measures for the *Bradford Univ.* and *Sheffield Univ.* are presented for their best run, for lack of detailed information in their reports. This dataset is very demanding, since it is composed by long videos with a high number of transitions within each one. Our results are very competitive. What is important to emphasize is the tradeoff of complexity and accuracy between our algorithm and the remaining ones that use sophisticated techniques based on features matching, self-supervision learning, among others, and involve particular modules to detect each type of transition and other artifacts, such as flash detection. This final evaluation gives evidence of the accuracy of the proposed algorithm, which makes it suitable to provide valid segments that summarize the video.

5.4.6 Video Summarization Impact

In this Section, we conduct experiments to evaluate the impact of our video summarization technique (see Section 5.3.3) in the overall classification performance. Namely, we use the *Hollywood2* dataset and consider the most simple approach where just the most similar shot, the first in our ranking, is selected and used to represent the whole video, defined here by *most-representative*

Method	P	R	F_1
Beijing Univ.	82.1	92.2	86.8
Bradford Univ.	**91.9**	**94.1**	**92.9**
Marburg Univ.	78.5	81.2	73.3
NHK	92.8	90.4	91.6
Sheffield Univ.	85.0	87.0	86.0
Our	85.8	90.9	88.1

Table 5.5: Results (%) in *TRECVID SBD 2007* dataset.

(*SBDMost*). This represents a preliminary test towards the chance of dramatically reduce the overall classification process, while keeping a competitive performance. From Table 5.6, we verify that performance drops by 2.4% and the computational time was reduced to 71.1%. This result is just an indicative of the potential of shot-based classification. However, more experiences should be done, specially in challenging videos, such as the ones presented in TRECVID MED, where camera motion is highly variable.

Method	mAP
Baseline	71.1
Baseline-SDBMost	68.7

Table 5.6: Results (%) in *Hollywood2* dataset considering the baseline, and the shot-approach *SBDMost*.

5.4.7 Contextual Features Impact

In this Section, we conduct experiments to evaluate the contribution of our proposed camera descriptor, D_{cam} in ***VIMUAR***. We observe in Table 5.7 a contribution over all the datasets, 2.1%, 1.1% and 1.7% in *Hollywood2*, *Olympic Sports*, and *HMDB51*, respectively.

From these results, specially for the highest contribution in the *Hollywood2* dataset, we may highlight the intuition traced previously in Section 5.4.3, that provides evidence about the robustness of our proposal in obtaining the global camera motion and modelling its behavior in the presence of close-up of actors and change of camera shots. Since D_{cam} includes complementary information from camera such as its parametric model, the radial projection response and the statistical behavior of surrounding cells in the frame, it will be interesting to decompose the descriptor and measure the contributions of each individual part. Another important experience will be to analyze in detail our sampling technique to acquire information about the variability and discriminative power of D_{cam}, for instance to distinguish the importance of the uniform sampling and the sparse sampling guided by the detection of camera type transition. With these results, we may consider that our intuitions for our individual contributions around camera motion stabilization and modelling gain some significance and pertinence for action classification in multimedia videos.

Method	Hollywood2	Olympic Sports	HMDB51
IDT-withHD (subset)	62.0	88.8	51.0
Baseline	71.1	89.9	50.4
Baseline + D_{cam}	73.2	91.0	52.1

Table 5.7: Impact of the proposed camera descriptor, D_{cam}, in the overall classification (%) of our baseline (mAP for *Hollywood2* and *Olympic Sports*, and mACC for *HMDB51*).

5.4.8 Motion Features Impact

This Section provides another way to inspect the contribution of our proposal for camera motion compensation. Table 5.8 shows the impact of the inclusion of the ω-DCS in ***VIMUAR***. Jain *et al.* [144] proposed this descriptor and they showed an extensive study about its impact in the overall classification performance. We have already discussed the most important conclusions in Section 5.3.4. Therefore, this Section aims to report an implicit comparison between their motion compensation solution and our proposal. They reported that the inclusion of ω-DCS improves the baseline between 0.9% and 1.2%. In ***VIMUAR***, we report a contribution from 0.4% to 1.4%, therefore presenting lowest and highest boundaries, but we also evaluate its impact in one more dataset, *Olympic Sports*. Further experiences, in more datasets, are needed to draw out a valid conclusion. We have already stated the contribution of our camera motion compensation in the baseline, therefore more detailed information should be extracted to infer the real impact of the warped flow in the DCS descriptor.

Method	Hollywood2	Olympic Sports	HMDB51
Jain *et al.* [144]	62.5	—	52.1
IDT-withHD (subset)	62.0	88.8	51.0
Baseline	71.1	89.9	50.4
Baseline + ω-DCS	72.2	90.3	51.8

Table 5.8: Impact of the ω-DCS descriptor as foreground component, in the overall classification (%) of our baseline (mAP for *Hollywood2* and *Olympic Sports*, and mACC for *HMDB51*).

5.4.9 Comparison with the State-of-the-Art

Here, we compare ***VIMUAR*** at its full potential, i.e. including the motion and contextual features, with recent methods in the literature. The results are reported in Table 5.9. From the three benchmark datasets, our method largely improves reported results in the state-of-the-art for one dataset, *Hollywood2*. However, we must highlight that we did not yet used the entire samples of any of the presented benchmark datasets, and as shown in Table 5.4.4, our proposal surpass the state-of-the-art work of Wang *et al.* [366], computed in the same subsets, in two datasets. It is mandatory to extract the results from the complete datasets for a fair comparison. However, we must revise the performance of ***VIMUAR*** in *HMDB51*. In general, the comparative results reveal promising performance for our proposal.

Method	Hollywood2	Olympic Sports	HMDB51
Jain *et al.* [144]	62.5	—	52.1
Jiang *et al.* [148]	59.5	80.6	40.7
Heilbron *et al.* [125]	64.1	92.5	57.9
IDT-withHD (from [366])	64.3	91.1	57.2
Peng *et al.* [273]	—	**93.8**	**66.8**
Oneata *et al.* [259]	63.3	89.0	54.8
Lan *et al.* [183]	68.0	91.4	65.1
Baseline + D_{cam} + ω-DCS	**73.8**	91.6	53.4

Table 5.9: Comparison with the state-of-the-art (%) of our baseline including D_{cam} and ω-DCS (mAP for *Hollywood2* and *Olympic Sports*, and mACC for *HMDB51*).

5.5 Summary

In this Chapter, the classification of multimedia videos by action recognition framework, ***VIMUAR***, is addressed in an innovative way from the literature. Many works have presented contributions for compensation of camera motion just to rectify or warp the image and optical flow. To the best of our knowledge, just one work [125] has explicitly included information of camera as contextual feature. However, from the experiments we state that our camera descriptor is more robust than their proposal. However, we believe that our camera descriptor is far away from its truly potential, since we did yet integrate all its features.

VIMUAR also includes an interesting approach for the detection of regions of interest, that permits to separate foreground from background regions. Such feature enables the improvement of the classification performance, while reducing, by a significant amount, the number of extracted trajectories. Separate experiments should be conducted to infer the real contributions of the camera motion compensation algorithm and of motion-saliency map.

The proposed video shot segmentation algorithm has shown competitive results when compared with more complex algorithms for the task of shot-detection in complex datasets. The video shot-summarization reveals promising results, since using just the most similar shot of the video, the final classification drops slightly. However, this result may not be conclusive since just one dataset was tested.

This approach encompasses the level of human action defined as *the whole*, through the complete proposal of ***VIMUAR***. However, further work need to be done to include a multimodal representation of the entire frame, not only by modelling the scene explicitly by a descriptor, but also through the inclusion of other representations of foreground motion.

The next Chapter changes the application domain to the Behavioral settings. It presents our approach to *the parts* level, in which the motion representation of upper body parts assumes a predominant role to infer different expressiveness behaviors in non-verbal communication, while considering the influence of emotional context.

Chapter 6

Body Expressiveness Analysis in Social Context*

> We are social creatures to the inmost centre of our being. The notion that one can begin anything at all from scratch, free from the past, or unindebted to others, could not conceivably be more wrong.
>
> Karl R. Popper

Body gestures serve as main communicative function and contain substantial affective and cognitive information that help us emphasize certain parts of our speech. As well, gestures are most of the times faithful to the speaker's communicative intention [117]. The interpersonal socio-emotional interaction and the recognition of a person's affective state are vital for communication. Concerning the computer vision field, the usage of automatic tools for the extraction of body features allows a better understanding of the human behavior. Decoding and studying expressiveness is useful for several applications, in which the analysis of behavioral video is included as a way to describe the social interactions between two people [306].

6.1 Introduction

Human activity analysis plays an important role in gathering meaningful information to understand human behavior in social relationships. Detecting and interpreting temporal patterns of nonverbal behavioral cues in a given context is a natural and often unconscious process for humans. However, this still remains a rather difficult task for computer vision systems [41]. The idea of reaching nonverbal sensitivity through computational models is very appealing since this may help us to understand how the human visual system interprets all the sensory events in the environment and

*Some portions of this Chapter appeared in [308]

how it relates those events. Such knowledge may be used to create applications that assist to mitigate social inclusion and social inequalities.

Previous chapters in this thesis have focused their study on either coarse motion representation, or relational features, or even abstract motion evidence in different settings. The current application settings of this Chapter involves the study of motion in human body parts to represent expressiveness and characterize its intention. It encompasses as context a duo-interaction between deaf and hearing people under several conversational topics with affective meaning. This behavioral scenario is part of a project that aims to study the Portuguese Sign Language (PSL) at two levels: i) recognition of the gestures involved at the sentence level of PSL; ii) expressiveness identification within social context. The ultimate goal of the project is to provide a digital mean to understand and reduce the communication gaps that isolate deaf people in several social activities.

Sign language speakers experience their language very passionately. This may be explained by the fact that language plays a crucial role in the construction of a community and it is a clear mark of belonging [237]. Emotion recognition from body language and its implications to the social adjustment of a sign language speaker are very important issues. Sign language expressions are composed of manual (hand gestures) and non-manual components (facial expressions, head motion, pose and body movements). Some expressions are performed only using hand gestures whereas some change the meaning when a facial and body expressions are included [19].

This Chapter presents a preliminary study to tackle the second level of the aforementioned project. It elaborates a motion-based analysis of body, namely the upper limbs, as a channel of expressiveness in nonverbal communication. Due to its novelty in terms of application and cross-domain concepts, the literature does not combine body expressiveness with sign language conversation within emotional context, to the best of our knowledge. Therefore a user-case scenario is defined, along with its variations and user needs. This case scenario does not correspond to the final application scenario, since it would represent the acquisition and mobilization of other resources not available at the present time. However, its methodology and needs are designed by a team specialized in the areas of PSL education and sociology to ensure the validity of the study. It focus on evaluating the differences between deaf and hearing people, in terms of expressive patterns through body gesture analysis, while considering the influence of social context, in terms of *conversational topics*. In more detail, the following problems are outlined: i) differentiation between deaf and hearing people; ii) identification of different *conversational topics* based on body expressiveness; iii) identification of different levels of mastery of PSL speakers through visual analysis. For such scenario, an unique video dataset, that encompasses socio-psychological premises, was recorded and partially annotated, an an expressiveness-based framework is proposed, named **VIdeo-based Body Expressiveness Analysis** (***VIBEA***).

6.2 Overview

6.2.1 Body Gesture

Gesture is the use of motion of the limbs or body as a means of expressing/communicating an intention or feeling [117]. This human communication channel is responsible for passing several signals to the context we are inserted in. Research on expression recognition in psychology and computer vision fields used to focus exclusively on photographs or video sequences of facial expressions, since this was considered to be the only powerful channel of nonverbal communication. More recently, researchers realized the amount of expressive information that body gestures may transmit in a more subtle way [22,288,362]. The recognition of human gesture has already shown its value in areas such as: behavior understanding, human-machine interaction, machine control, surveillance, among others [139].

According to Cassell [58], the fact that *gestural errors* are extremely rare demonstrates how essential their nature is for accurate communication. If we think about spoken language, we can realize how disfluent it can be, full of false starts, hesitations, and speech errors. On the other hand, gestures are almost always faithful to the speaker's communicative intention. Five different types of gestures may be defined [117]: i) *Gesticulation*, movements of the hands and arms that accompany speech; ii) *Language-like gestures*, gesticulation with the intention of replacing a particular spoken word or phrase; iii) *Pantomimes*, gestures that depict objects or actions, with or without accompanying speech; iv) *Emblems*, familiar gestures such as, thumbs up, and assorted rude gestures (often culturally specific); v) *Sign languages*, linguistic systems which are well defined.

Gesture recognition, a relevant cue to be analyzed in the case of sign language, can be defined as the process by which a human observer or a machine identifies the body gestures made by an individual. If we think about which types of gestures would be more easily recognized by computer systems, we would come to the conclusion those would be emblems or even sign language since they tend to be less ambiguous, less natural and more likely to be learned. Emblematic gestures carry more clear semantic meaning. Several steps can be identified that are transversal to most gesture recognition systems [117]:

1) Sensing human position, configuration, and movement in the scene using cameras and computer vision techniques;

2) Preprocessing;

3) Gesture modelling and representation;

4) Feature extraction and gesture analysis;

5) Gesture recognition and classification.

Video-based approaches for gesture recognition are challenging due to spatio-temporal variations and endpoint localization issues. An effective gesture recognition system is introduced by Ravindra de Silva *et al.* [79], in which Discriminant Analysis is used to build affective posture predictive models and to measure the saliency of the proposed set of posture features in discriminating between four basic states: angry, fear, happy, and sad. Li and Greenspan [193] also work on

segmentation and recognition of dynamic gestures reaching an effective model based on Dynamic Programming. In the work proposed by Piana *et al.* [288] anger, disgust, fear, happiness, sadness and surprise are six states evaluated through the extraction of features such as kinectic energy, quantity of motion, barycenter tracking, etc. This work is performed in a controlled environment to monitor the behavior of autistic children while playing with an interactive serious game.

In conclusion, gestures serve an important communicative function in face-to-face communication since they often occur in conjunction with speech. Studies have also shown that, even in the absence of facial and vocal cues, it is possible to identify basic emotions signaled by static body postures [21], arm movement [288] and whole body movement [21, 239, 288].

6.2.2 Expressive Gesture

The concept of gesture leads us to think about a set of temporal-body features responsible for conveying expressiveness. The scientific community is gaining interest in studying the ways of modelling and communicating expressive content in non-verbal interaction. Two performing arts in which this increased interest is evident are: the study of movements in music performances to evaluate the conductor, and full-body movements study to evaluate the performance of a dancer [55]. In this particular context, gesture is considered to contain and convey information related to the emotional-affective domain, which is different from the traditional meaning we give to language gestures. In this context, gesture is the responsible for conveying expressive content, i.e., relevant information that is suggestive of a certain state of emotion [55].

For the particular purposes of our work, it is important to assimilate that gestures and the way they are performed work as an identity of an individual. The same action may be executed in several different ways depending on the executant. It is possible to infer emotional states by the way a person walks, so-called gait analysis [174]. This perspective permits to classify walking as an expressive gesture [139]. Several everyday actions may constitute expressive gesture: Pollick *et al.* [291] investigate expressive content of actions such as knocking or drinking, Heloir and Gibet [127] work on the identification and representation of the variations induced by style for the synthesis of realistic and convincing expressive gesture sequences in sign language speakers.

It is possible to identify the main challenges concerning the study of expressive gestures [319]:

i) Define and characterize the expressiveness and variability in human movement. This expressiveness is considered at all levels of gesture generation, and involves both a semantic dimension (from actions that convey a specific meaning to sign languages that imply the linguistic aspects of phonetics, phonology, prosody, etc.), and an expressive dimension induced by intentional variations or emotional states of the actor, and results in variations in the produced signals;

ii) Explore new motion representation spaces that reflect the expressiveness and variability contained in the data. This implies reducing the complexity of the high-dimensional motion data by proposing different embeddings for these data. Such embeddings should enable to characterize

and parameterize specific action sequences, which give rise to original approaches for recognition, or generation of new behaviors, inspired by sensorimotor biological processes;

iii) Link the different levels of representation, from narrative scenarios through structural patterns of actions, to continuous streams of motion data. More precisely, the aim is to extract structural patterns from data and to understand how these discrete patterns influence the synthesis of gestures while preserving the semantics of actions as well as subtle expressive variations;

iv) Defining evaluation protocols that are necessary for evaluating the different hypothesis and models that are constructed at all the levels of the perception-production loop.

Automatic analysis and synthesis of expressive gesture can open novel scenarios in the field of interactive multimedia. Computational models of expressiveness in human gestures can contribute to new paradigms for the design of interactive systems, improved presence and physicality in the interaction.

6.2.3 Body Expressiveness Representations

In order to be successful interpreting the expressiveness of a human movement/gesture automatically, it is crucial to use precise models that can capture that expressiveness. When performing this task, it is important to consider not only the activity of interest, but also the whole shared environment in which the activity is impregnated in, scene and context. Thus, modeling the activity is concerned not only with modeling the action performed by different objects in isolation, but also the interactions and causal relationships among these actions [387]. The fusion of several elements such as motion, appearance, shape, among others, is the cue for solving this fundamental, yet challenging, research topic that has driven the efforts of many researchers. Two types of behavior representation strategies are addressed next: trajectory-based and pixel-based.

6.2.3.1 Trajectory-based Representations

A trajectory representation is a type of object-based representation which consists of describing in detail objects in a space over time based on the assumption that individual objects can be segmented reasonably well in a visual scene [**?**]. Trajectory-based representation is a method implemented for systems using passive or active markers which yield a high contrast in the images and provide a robust representation [227]. It aims to construct spatio-temporal trajectories centred on each object's boundary box or shape structure. The object trajectory is the history of the motion in a visual scene [**?**]. A trajectory of a certain object is computed by associating that object in consecutive frames using motion tracking [365]. To achieve the tracking step, it is necessary to apprehend some object appearance attributes like colour, shape or texture so that it is possible to establish inter-frame correspondence over time. In a trajectory-based representation a single trajectory track should be associated to an object of interest.

Unfortunately, image quality issues make this task hard, for instance due to noise and crowded environments. Some techniques have been developed to cope with occlusions and lighting changes [121, 217], and by adopting a tracking-by-detection strategy [45, 380]. Despite these scientific

efforts, problems remain unsolved on how to best combine useful information sources and filter out unreliable information, so that object tracking is not lost over time and that false positives do not occur.

Many human action recognition studies benefit from this approach [149, 256, 318] using it as a starting point for behavior analysis. The main problems encountered by the authors of these studies, that in some ways compromise the results, are: i) Lack of details: a cluttered public scene may have low fidelity visual detail, which prevent from extracting sufficient image features; ii) Severe occlusion: lost of the points/regions that are being tracking over time may happen if the objects are inserted in a large space shared with other objects where inter-object occlusion may occur. This leads discontinuous trajectories and inconsistent labelling of objects; iii) Lack of context: an isolated trajectory of an object does not always capture distinctively the information that one may be looking for. Interpreting the behavior of an object in an unconstrained environment analysing only its trajectories appears to be insufficient in several cases.

6.2.3.2 Pixel-based Representations

A pixel-based strategy neglects individual object entities, targeting its focus to all relevant pixels found on the image, colour and intensity gradient information. It is extremely useful since it is computationally simple and appears to be beneficial for representing object activities in cluttered or crowded scenes. It is most commonly used when an exact distinction between what constitutes foreground and background is wanted, for example in a busy urban scene with different elements such as people (foreground) and buildings, trees and permanent objects (background) [314]. Effective techniques have been implemented from a short block of input frames by casting the problem as an exercise for optimal labeling. These are useful approaches of background approximation that assign a set of labels or pointers for each pixel in the background image, such as: Combinatorial Optimization, Minimum Cut/Maximum Flow and Alpha Expansion [67].

Although this pixel-based approach is considered a simple one, it can be of much value for modeling and understanding actions and activities [**?**]. After foreground pixel detection, using for example an adaptive Gaussian mixture background model [340], more complete pixel-based representations may be obtained such as for example:

i) The Motion History Image (MHI), is used to detect visual changes in images by keeping a history of the changes that decays over time [38]. It is a widely used algorithm due to its simple implementation and easy visualization. Looking at Fig. 6.1 it is intuitive to perceive several aspects such as the direction of the motion, the current position of the object moving and its past localizations on the scenario. For instance, Babu and Ramakrishnan [25] include MHI representation, among others, into a feature-based statistical framework for gesture recognition.

ii) Pixel Signal Energy (PSE), uses temporal filters to measure the average magnitude of pixel-wise temporal energy over a backward window, where the size determines the number of frames needed to be stored [243]. It allows to extract reliable temporal changes at individual pixels. Furthermore, pixel energy constitutes a good measure for exploiting synchrony in pixel-events. For instance, Ng and Gong [243] address the limitation of the short-sighted view of single pixels

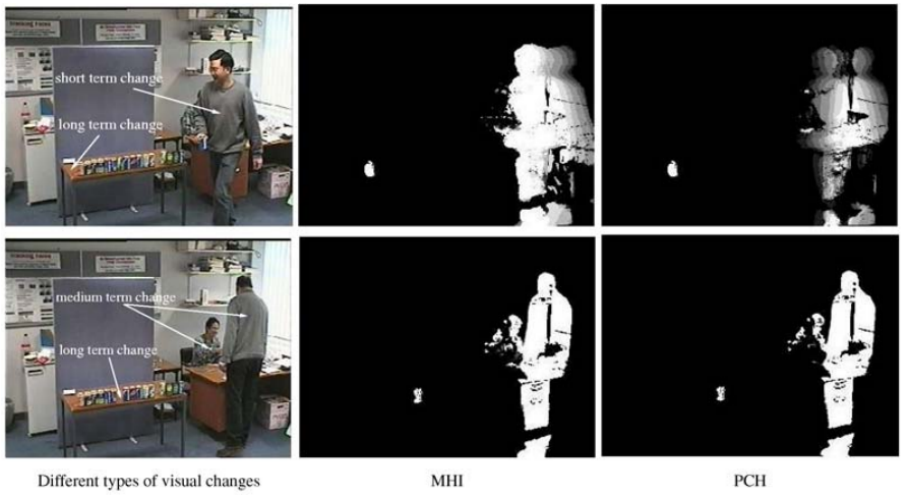


Figure 6.1: Examples of Motion History Images and Pixel Change History images for different types of visual changes. Extracted from [388].

providing a more flexible framework for capturing global events as opposed to related spatio-temporal motion-energy measures.

iii) Pixel Change History (PCH), combines the two previous methods measuring the multi-scale temporal changes at each pixel [389]. A PCH image becomes a MHI image if its accumulation factor is set to one. This method can capture a zero-order pixel-level change, i.e., the mean magnitude of change over time [388].

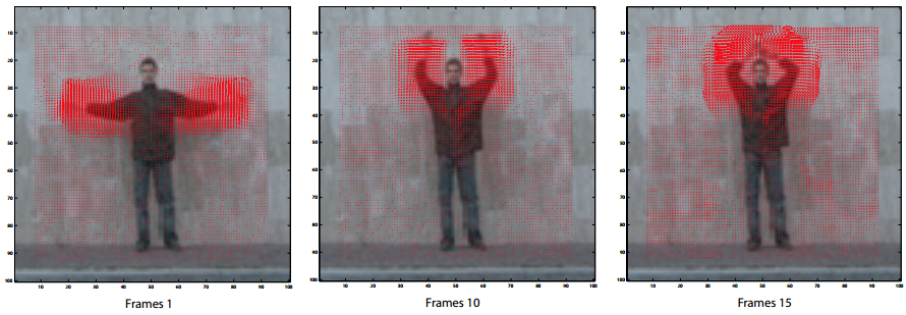


Figure 6.2: Optical flow of arm-waving action at frames 1, 10 and 15. Extracted from [14].

If the assessment of an object's spatio-temporal information, particularly facial and body features motion, is intended, many researchers use optical-flow techniques [14, 104, 249, 397]. An optical flow field is usually a dense map of imagery displacement vectors that estimate pixel-wise apparent motion between consecutive image frames over time [**?**]. When it comes to facial and body expression recognition for example, a sequence of images contains much more information than a single image as it is visible on Fig.6.2, where an example of source vectors is shown for the detection of the wave of the individual's arms [14]. Flow vectors are computed by image grids instead of per pixel which happens to be computationally expensive. Thus, similarly to foreground pixel-based features, the optical flow-based representation avoids tracking individual objects.

6.3 Dataset

Given the specific nature of the defined user-case scenario, the construction of a dataset is mandatory. The literature presents several datasets for facial expression analysis at different levels and for different tasks, such as Cohn-Kanade [1], MMI [2], EURECOM [3], Bosphorus [4], among many others; and others also integrate body gesture analysis such as FABO [5]. Regarding the topic of automatic sign language recognition, several datasets were already introduced in the literature, like the American Sign Language Lexicon Video Dataset (ASLLVD) [6], RWTH-BOSTON-10-50-104-400 [7], Corpus NGT [8], British Sign Language Corpus Project (BSLCP) [9], among others. However, there is no dataset in the literature that combines body and facial expressiveness analysis for sign language within several emotional context.

One of the main advantages of the construction of this dataset is its richness in terms of behavior-related features. Researchers like the ones that developed the studies discussed so far, which are inserted on the context of computer vision for expressiveness and emotion analysis, would certainly benefit from a dataset containing this type of information regarding the evaluation of body gestures, their expressiveness and contextual variations. Several state-of-the-art studies are focused on the extraction of expressive cues from human gesture, aiming to prove that these are valuable indicators of a person's profile and emotional state. In this sense, this dataset intends to serve in the future as a support to extract information about the social interaction of two persons, in this case deaf and hearing, in order to infer if there are social difficulties in terms of communication and detect different levels of empathy.

6.3.1 Definition

The dataset contains people from the two populations, deaf and hearing people. Its motivation is to enable studies for the analysis of dialogue relationships between two individuals in a relaxed environment from different perspectives. Its design takes in consideration not only technical issues, but also sociological and psychological considerations. Its preparation and acquisition process was carefully advised and discussed with staff of the lab of social-psychology of the University of Porto[1]. It worths to mention that from a socio-psychological point of view, other type of assumptions and a deeper study would be necessary for a fully validation in such area.

In order to enable the conditions to study the proposed research questions stated in Section 6.1, different aspects were considered for the acquisition process. For instance, the sample set is

[1] http://www.pitt.edu/~emotion/ck-spread.htm
[2] http://mmifacedb.eu
[3] http://rgb-d.eurecom.fr
[4] http://bosphorus.ee.boun.edu.tr/default.aspx
[5] http://www.eecs.qmul.ac.uk/~hatice/fabo.html
[6] http://www.bu.edu/av/asllrp/dai-asllvd.html
[7] http://www-i6.informatik.rwth-aachen.de/aslr/index.php
[8] http://www.ru.nl/corpusngten/
[9] http://www.bslcorpusproject.org
[1] Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto - http://sigarra.up.pt/fpceup

important and transversal to all the experiments. The available sample set was constrained by the available resources of our partnership, namely the Agrupamento de Escolas Eugénio de Andrade, Escola EB2/3 de Paranhos. Following the advise of the students from the lab of social-psychology, the samples were organized in a balance manner considering several aspects such as gender, age, nationality, education level, occupation, current employee status, LGP expertise level, and level of familiarity with the deaf community. Namely, the following main constraints were defined:

i) The sample set should be equally composed by the same number of deaf and hearing (sign language speakers) people. For convenience, six hearing and six deaf people were considerated;

ii) The individuals should know each other *a priori* and have some kind of affinity;

iii) The acquisition should take place in a venue that is familiar to all individuals.

The first constraint controls the variables *gender* and *age*, and it represents a sample set called *by convenience*, regarding the psychology terminology. The remaining constraints aim to reduce the level of shyness and discomfort between the individuals, promoting a more engaging and natural conversation.

In order to be able to differentiate between deaf and hearing people, the population is equally subdivided among both groups (six deaf and seven hearing individuals) and different *scenarios* were considered (see Table 6.1). The sample set includes school elements who participated voluntarily in the creation of the dataset. Table 6.2 shown its characteristics by group (deaf or hearing), gender and age. Refer to Section 6.3.2 for a description of the sample set.

Scenario	Description
S1	conversation between two deaf people.
S2	conversation between two hearing people.
S3	conversation between a deaf and a hearing person.

Table 6.1: Designed *scenarios* for the differentiation among deaf and hearing people.

Group	ID	Gender	Age
Deaf People	D1	Female	38
	D2	Female	35
	D3	Female	36
	D4	Female	39
	D5	Female	31
	D6	Female	39
Hearing People	H1	Female	38
	H2	Female	27
	H3	Female	30
	H4	Female	29
	H5	Female	30
	H6	Female	37
	H7	Female	37

Table 6.2: Sample set for the creation of the dataset. All individuals are woman between the ages of 27 and 39.

Taking in consideration the conceived *scenarios* (see Table 6.1), the individuals are coupled so that all the conversation *scenarios* are covered. Each conversation between a pair of individuals is designated as a *session*, with nine sessions included so far in the dataset, three for each *scenario*, as indicated in Table 6.3. An individual can be part of more than one scenario depending on his/her availability.

Session	Pair	Scenario
01	D3 & D2	S1
02	H3 & H5	S2
03	H1 & H4	S2
04	D1 & H6	S3
05	H6 & H2	S2
06	D6 & H2	S3
07	D5 & D6	S1
08	D6 & D4	S1
09	D5 & H7	S3

Table 6.3: *Sessions* comprising the different pairs of individuals featured in the dataset.

Regarding the research question that ambitions to distinguish the body expressiveness in several contexts, different *conversational topics* were conceived to awake certain emotions in the individuals. Those topics are defined in a staggered way, so that the discussion would generate emotions of increasing intensity in the individuals of the conversation. The topics are chosen assuming that a dialogue would occur between a pair of individuals and that both would intervene actively. Four different *conversational topics* are defined belonging to two-fold *emotional moments*: positive (T1 and T2) and negative (T3 and T4). The topics should be discussed by the order stated in Table 6.4.

The dataset is composed by 36 video-sessions (nine *sessions* and four *conversational topics*), since each *session*, i.e. each pair of individuals defined by the three *scenarios*, has four *conversational topics*. Fig. 6.3 illustrates the structure of the dataset in terms of content.

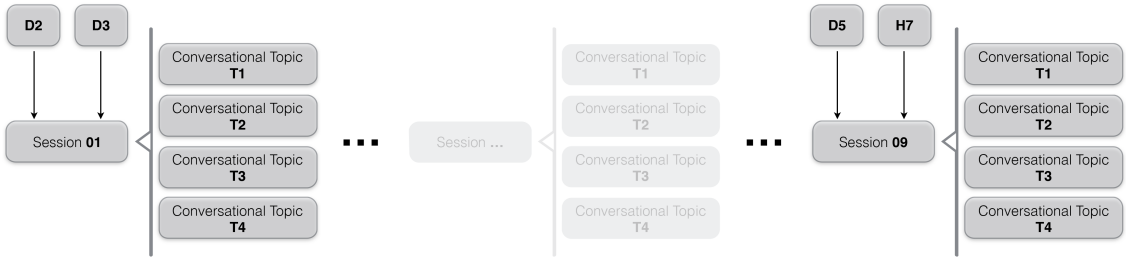


Figure 6.3: Diagram of the dataset structure, showing the individuals that form a *session* and the four *conversational topics* for each *session*, summing a total of 36 video-sessions.

In order to acquire relevant information to help to infer the different levels of mastery of PSL speakers, all the individuals involved in the dataset were inquired through a formulary, which takes into consideration the advises of PSL instructors from the Escola EB2/3 de Paranhos and the

T1 - Positive: Talk about happy moments
• Describe several happy moments that happened throughout the person's life.
• Explain why those moments were happy.
• Explain their context and circumstance.
• Mention and explain some happy moments outside of the actor's personal life.
T2 - Positive: Talk about people with which the actor has a strong love or friendship bond
• Describe those people.
• Explain the reason why the strong bond exists.
• Explain the importance of those people in a personal point of view.
T3 - Negative: Talk about sad moments
• Describe several sad moments that happened throughout the person's life.
• Explain why those moments were sad.
• Explain their context and circumstance.
• Mention and explain some sad moments outside of the actor's personal life.
T4 - Negative: Talk about situations that awaken anger/indignation/injustice
• Explain why those situations cause anger/indignation/injustice.
• Explain how those situation may affect personal life on the short and long run.
• Explain their context and circumstance.
• Mention a few typical situations of that kind.

Table 6.4: *Conversational topics* defined to distinguish the body expressiveness in several contexts.

staff of the social-psychology lab. Two versions of the questionnaire were used, one for the deaf individuals and another for the hearing ones (Fig. A.1 and Fig. A.2 in the Appendix).

6.3.2 Sample Set Description

Our sample set is composed by 13 women of portuguese nationality, six individuals from the deaf community and seven from the hearing community, between 27 to 39 years of age (μ=34.5, σ=4.3). From the sample set, 92.3% (n=12) of the individuals have a higher education, being that all of them currently carry out a professional activity in the area of PSL education. All of the individuals are familiar with the deafness, since they have deaf friends and/or deaf family members. All of them also belong or keep contact with private/public educational entities for/with deaf people. The parents of the deaf people are hearing. Only one of the deaf people was born deaf. From the deaf people, two of them use hearing aids and the remaining have use them in the past.

6.3.3 Support Material

All the individuals filled a questionnaire at the end of the acquisitions. The research team informed the individuals that the questionnaire would be used as a support material for the study, namely for its socio-demographic characterization. The participation was not mandatory and guarantee of confidentiality was assured. The questionnaire also includes an opinion-based scale, from 1 to 9 (1=completely disagree, 9=completely agree), which was used to inquiry their opinion about the current social situation of the deaf community in Portugal.

6.3.4 Physical Setup and Technical Characteristics

The Escola EB2/3 de Paranhos was the selected venue for the acquisition to take place because of the mentioned partnership and also because it was a familiar environment for the volunteer population. A large classing room was equipped with four IP cameras (IP0, IP1, IP2 and IP3) and one Microsoft Kinect (K). Fig. 6.4 represents the acquisition layout. These were all placed in strategic locations (at a height of 2.58 meters) for the best possible capture. The location of the cameras was hidden from the individuals so that the dialogues and reactions would not be affected by the awareness of the presence of the cameras, or at least minimized. Two chairs were centred in the room in a way that was propitious for the dialogue in terms of proximity and comfort, and for the video acquisition. All distances are presented in meters. Observing the Fig. 6.5, it is possible to visually understand the perspective captured by each of the cameras at the same time instant. The conversation *scenarios* were recorded in a video format and some statistics organized by *session* are presented in Table 6.5.

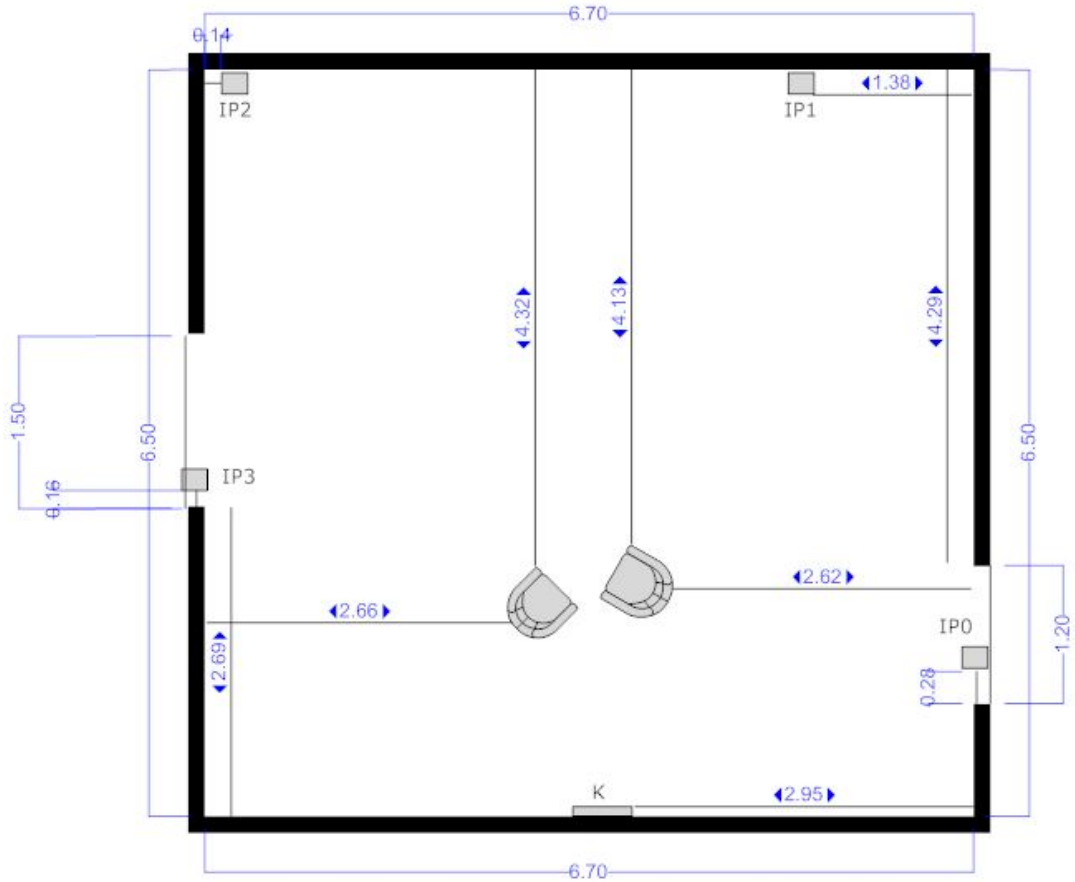


Figure 6.4: Sketch of the dataset acquisition set. The four different cameras used are IP0, IP1, IP2 and IP3. A Microsoft Kinect is represented with a K.

Figure 6.5: A frame of a dialogue moment captured from the different cameras P0, P1, P2 and P3. The same time frame is displayed from the perspective of the four cameras.

Session	Duration (s)	# Frames	# Videos
01	136	1934	4
02	495	7856	4
03	338	5166	4
04	415	6935	4
05	274	6837	4
06	273	6803	4
07	255	4532	4
08	196	2496	4
09	496	6295	4

Table 6.5: Dataset statistics, in terms of duration and length, aggregated by session. Average results are shown. The last column represents the number of videos per camera.

6.3.5 Protocol

The acquisition protocol is very simple, since all the preparation was previously executed as described in Section 6.3.1 and Section 6.3.4, including the definition of the pairs. The only information that was given to the individuals was that they would participate in a study about PSL. Any information about cameras, recording, analysis of expressiveness, or related information was intentionally omitted by the research team.

Prior to the start of the acquisitions, the individuals had time to read a script with the topics they would have to talk about, so that they could prepare and remember about some of the situations that they are asked about. Upon completion of each *conversational topic*, the pair of individuals leave the room, close the door, and re-enter to start the next moment. At the end of the acquisitions, all the individuals filled the questionnaire in which demographic and opinion enquiries were included (see Section 6.3.3).

6.3.6 Data Selection and Preparation

Taking into account the aforementioned research questions, and considering that in the conversations between two hearing people sign language is not used, such sessions are not considered in this work. The subset of the dataset used for building the framework of this study is represented by the sessions 04, 06, 07, 08 and 09. Session 01, featuring individuals D3 and D2, is also discarded since after post-visualization of the recording, we notice that the position of the individuals on the set was different, they were standing, while in the remaining sessions all the individuals were sitting. Therefore, it would not be possible to use those individuals as a comparison with the other individuals.

The analysis of the expressiveness of body gestures is performed in a frontal view, since it is the most valuable viewpoint for this kind of analysis. In this way, the data considered are from cameras IP1 and IP2, where each one records a frontal view of a single individual. Summarizing, the subset of the dataset is composed by seven individuals (four deaf and three hearing people), two *scenarios* (deaf-deaf and deaf-hearing), five *sessions* (04, 06, 07, 08 and 09), four *conversational topics* (T1, T2, T3 and T4), for a total of 20 video-sessions (40 videos, since each session has two individuals, and just one camera is used for each one).

For each video, the following statistics were measured: frames per second (fps), total number of frames and duration. Each one was subdivided into *mini-clips* with the same number of frames and the same fps to be used as samples for further processing, learning and classification tasks. Depending on each video's fps, some videos underwent a downsampling and others an upsampling process. The final number of samples (*mini-clips*) are described in Table 6.6. It worths to mention that the Microsoft Kinect data was not used in this specific line of work.

6.4 Body Expressiveness Analysis Framework

6.4.1 General View

Emotion recognition from body language and its implications to the social adjustment of a sign language speaker are very important issues. ***VIBEA*** aims to provide solutions for the problems specified in Section 6.1, and its diagram is illustrated in Fig. 6.6. It presents two operational flows: i) *classification approach*, which identify and evaluate the differences between deaf and hearing people, in terms of expressive patterns through body motion analysis, considering, and also classifying, several emotional context, so-called here *conversational topics*; ii) *clustering*

Session	Individuals	Topic	# of *Mini-clips*
04	H6	T1	33
		T2	43
		T3	62
		T4	44
	D1	T1	33
		T2	42
		T3	60
		T4	33
06	H2	T1	32
		T2	20
		T3	42
		T4	35
	D6	T1	31
		T2	19
		T3	42
		T4	35
07	D5	T1	26
		T2	22
		T3	35
		T4	37
	D6	T1	26
		T2	21
		T3	34
		T4	37
08	D4	T1	21
		T2	10
		T3	19
		T4	23
	D6	T1	19
		T2	10
		T3	19
		T4	37
09	D5	T1	25
		T2	58
		T3	53
		T4	89
	H7	T1	26
		T2	58
		T3	53
		T4	64
Total			1266

Table 6.6: Number of *mini-clips* per individual in each session and for each *conversational topic*.

approach, which aims to find a hierarchical structure regarding different levels of mastery of PSL and organize the individuals within.

VIBEA processes each *mini-clip* frame-by-frame and extracts temporally different motion features, some of them appearance-based and other kinematic-based as detailed in the next Section 6.4.2, through the video. To obtain the final feature vector that represents the current *mini-clip*,

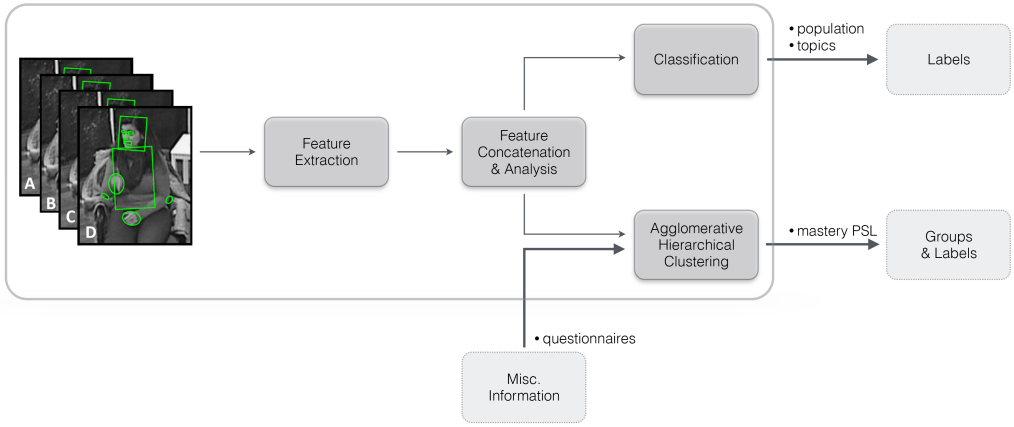


Figure 6.6: ***VIBEA*** framework for body expressiveness analysis.

the extracted features are concatenated, and then may go through a feature selection analysis and a dimensionality reduction process. Finally, the learning step point towards a supervised classification approach, or follow an unsupervised clustering mechanism guided by external information, such as variables gathered from questionnaires.

6.4.2 Feature Construction

As stated in Section 6.3.6, the *mini-clips* are used as samples in ***VIBEA***. For each one, a region of interest (ROI) around the area bounding the individual and his/her gestures is manually defined and considered for further processing.

Finding a suitable data representation is not an easy task being strongly related to the measurements that are possible to perform with the available data [119]. We explore several trajectory and pixel-based features to capture and represent body expressiveness regarding our goals. Some of those features are MHI, motion gradients, several body part trackers, and other kinematic features, which are described next.

6.4.2.1 Motion History Image

The MHI is a static image template where the pixel intensity is a function of the recency of the motion in a sequence. Basically in an MHI, pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. One of the advantages of the MHI representation is that a wide range of temporal motion events may be encoded in a single frame, and in this way, the MHI spans the time scale of human gestures. The MHI $H_\tau\ (x,y,t)$ can be computed from an update function $\psi(x,y,t)$ [9]

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \psi(x,y,t) = 1 \\ max(0, H_\tau(x,y,t-1) - \delta) & otherwise \end{cases} \tag{6.1}$$

where x, y and t are the position and time, respectively, $\psi(x,y,t)$ signals object's presence (or motion) in the current video image, the duration τ decides the temporal extent of the movement (e.g., in terms of frames), and δ is the decay parameter.

The value to be included in the final feature vector that will characterize each *mini-clip* is the mean value of the Quantity of Motion (QoM) in each frame, being the QoM the sum of all the pixel's intensities of the MHI: a *mini-clip* in which wider movements are performed will have a higher QoM. The standard deviation of the QoM of all the frames of a certain *mini-clip* also contributes for the final feature vector. The *mini-clips* in which the mean QoM was too low are suppressed.

6.4.2.2 Motion Gradient

The motion gradient (MG) is a feature that is obtained by the computation of the gradient of the MHI. Using this feature it is possible to get direction vectors pointing in the direction of the movement of a silhouette. The computation of this gradient is obtained by performing the convolution with separate Sobel filters in the X and Y directions [43]. Mathematically, the Sobel filter uses two 3×3 matrices that are convoluted with the image to compute approximations of the image derivative. One matrix is used for horizontal variations and another for vertical [43], as stated by

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{6.2}$$

and the orientation is then computed in the following way

$$orientation(\phi) = arctan\left(\frac{\frac{\partial MHI}{\partial y}}{\frac{\partial MHI}{\partial x}}\right) \tag{6.3}$$

The value to be included in the feature vector is also computed weighting for each frame the orientation obtained on each pixel by the value of the MHI on that pixel, giving more importance to the orientation associated to the pixels with a more recent motion. Mean and standard deviation values of the weighted orientation are included into the final feature vector of each *mini-clip*.

6.4.2.3 Motiongrams

One of the features explained previously was the QoM. It is a rough estimation of the amount of movement, and does not tell anything about the location of the movement inside the frame. This makes it difficult to know where the movement happened in the frame, for example whether movement was happening in the head, torso or feet. A solution to complement this idea would be to create a display that could visualize the QoM as well as the distribution of QoM in time and space. We propose to combine motiongrams with MHI, in order to overcome these problems.

The idea of the motiongram representation emerged from an analogy with the widely used waveform displays and spectrograms for efficient visualisation of some important features of audio material. The motiongram is an approach in which motion is measured by summing up the active pixels in a motion image and plotting the value over time. The motiongram image obtained gives the notion of the overall motion qualities in a video, while preserving some of the spatial information of where in the image/body the motion has occurred [147].

For our *mini-clips*, two types of motiongram are obtained: a vertical one, for which the MHI intensity over all rows in a column is assessed, and a horizontal one, for which the MHI intensity over all columns in a row is assessed. The motiongram approach is based on collapsing a matrix of size $M \times N$ into two different ones of size $M \times F$, in which M is the number of rows and F the number of frames in a *mini-clip*, and $N \times F$, in which N is the number of rows and F the number of frames, for horizontal and vertical motiongrams, respectively. Fig. 6.7 shows an overview of the extraction of a motiongram.

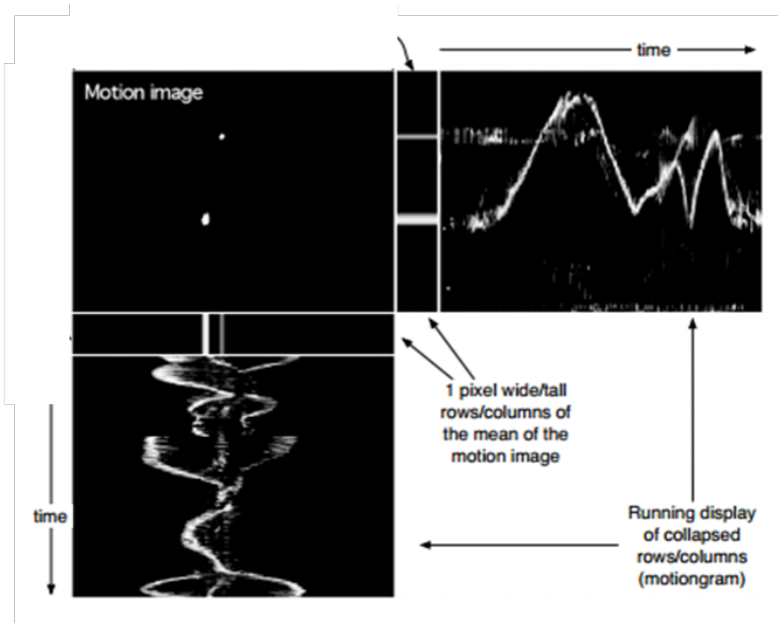


Figure 6.7: An overview of the process of creating a motiongram, showing the motion image, and the running motiongrams. Extracted from [147].

The features used for each *mini-clip* are histograms in which every bin contains the normalized sum of the motiongram information over all frames for each column and row for vertical and horizontal motiongrams, respectively. Special care is needed to guarantee that all the feature vectors for each *mini-clip* have the same dimension.

6.4.2.4 Body Part Tracking

Since the motion of different body parts represents a valuable insight for the study of motion patterns detection and expressiveness analysis in sign language gestures, some preliminary annotation methodology was conducted for this line of work. We did not included this information as part of

the dataset, since just a small portion of the entire dataset was annotated and used as guidance for the automatic tracking process described next.

The annotation took into consideration that the annotated spatial attributes should be as similar as possible, in terms of shape, with the body part that they represent. The annotation methodology also aggregated facial parts, but they were not used for this line of work. In sum, the annotated body parts so far are: i) Head, nose, mouth and trunk, represented by oriented rectangles; ii) Elbows, eyes and hands, represented by ellipses. The selected body parts for the extraction of the kinematic features are the *head*, *right arm* and *left arm*. The arm is considered as the enclosing rectangle that joint the elbow with the hand. Each body part has a representative point that correspond to the most outward point within that region, measured relatively to the body center of mass. In this way, for the *head* it is the top of the head, and for the *arms* they are the middle finger of each hand. To annotate each *mini-clip*, an uniform temporal sampling (on average 15 frames per *mini-clip*) was considered, where the initial frame was always annotated.

Fig. 6.8 shows the diagram of operations required to extract the kinematic features from each selected body part. The tracking step is never an easy task having most of the times to be executed in a semi-automatic way. In this case, for each body part the Lucas-Kanade feature tracker [205] is initialized with the respective representative point. A distance threshold is defined so that if the matching distance is exceeded, the tracking could be restarted with a new point manually inserted. The tracking process is automatically corrected on every frame there is an annotation. With the tracking information from the points that represent each body part, it is possible to extract the following relevant features:

i) *Occupation Rate*, using a grid that divides the image into cells, the percentage of occupation in each cell of a certain point being tracked is updated over time (for the whole duration of the *mini-clip*). The extracted features are the mean and standard deviation occupation rate from all the cells of the *mini-clip*.

ii) *Kinematic Features*, from the tracked points, referred above, it is possible to extract the respective trajectories and some kinematic features for both X and Y components, such as velocity, acceleration and movement direction. The mean and standard deviation values are calculated for each *mini-clip* and added to the final feature vector.

6.4.3 Distinguishing Deaf from Hearing People

The task of distinguishing between deaf and hearing people is one of the research questions of this work. Its relevance comes from the need to identify if a deaf person is well integrated into the current social context, or if he/she is being discriminated. Technically, this is a classification problem where the class are known *a priori*, namely if the individual is a deaf or a hearing person. Therefore, it follows the *classification approach* under ***VIBEA*** (see Section 6.4.1).

For classification purposes, the *mini-clips* are grouped by topic so that it is possible to compare the classification performance for the different groups. Table 6.7 shows the samples in which the *mini-clips* are grouped. Two classification methods are considered: *k*-Nearest Neighbours (*k-nn*), which is simple and widely used as a first approach for classification problems, and Support

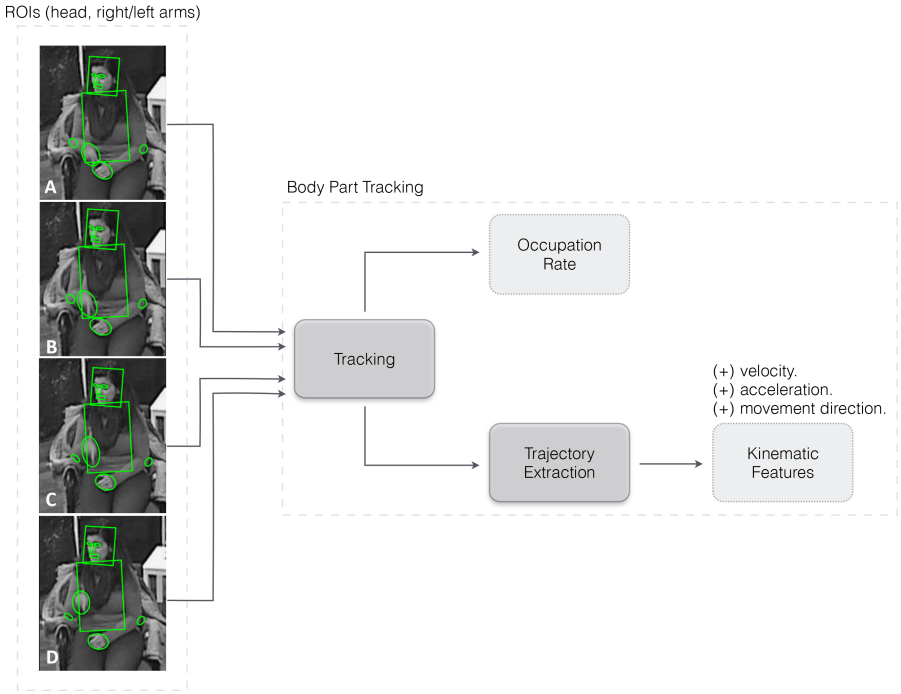


Figure 6.8: Diagram to extract the relevant features from different body parts.

Vector Machine (SVM), which is usually more accurate. Both algorithms are used under a cross-validation mechanism with different number of folds, dependent on the number of samples per group, to avoid over-fitting. For the SVM, we perform a grid search to automatically infer the optimal parameters. The statistical measures used to evaluate the performance of these two classifiers are: Correct Rate (CR), Recall (R) and Precision (P). The Confusion Matrix (CM) is also shown, sometimes accompanied by the worst CM along the classification process.

Group	# Samples
T1	234
T2	296
T3	361
T4	375
All	1266

Table 6.7: Sample groups for classification of the individuals.

6.4.4 Distinguishing Different Conversational Topics

Regarding the differentiation of the *conversational topics*, recalling that there are two with positive emotional connotation and two with negative, spotting the differences in terms of expressiveness among the four topics is the primary reason why these moments are included on the dataset. Its relevance comes from the need to identify the context to adapt the way an individual is being

monitored or adapt the behavioral monitoring conditions of the individual. This is also a classification problem. In this case, the *mini-clips* are grouped by individual, as shown in Table 6.8, and there are four classes to classify. In the same way as previously, the *classification approach* under ***VIBEA*** is adopted (see Section 6.4.1).

Group	# Samples
D1	158
D4	73
D5	328
D6	265
H2	97
H6	157
H7	188
All	1266

Table 6.8: Sample groups for classification of the *conversational topics*.

The same classifiers, *k-nn* and SVM, are used to approach this problem. However, we use a multi-class SVM for ordinal data, since we want to inspect if the topics follow a natural order of expressiveness, and analyze their intra-class and inter-class boundary decision between the topics that belong to the same positive or negative connotation. We used the approach generalized by Da Costa *et al.* [289]. The same metrics are used for evaluation performance.

6.4.5 Identifying Levels of Mastery in PSL

The purpose of distinguishing different levels of expertise in PSL is very abstract, but also represents a valuable addition to ***VIBEA***. Its relevance comes from the need to identify if the level of mastery of PSL could be a cause for social inclusion among the deaf people. This context involves a high degree of uncertainty regarding if the information extracted from the *mini-clips* insures a correct support to answer this research question.

Since the level of mastery of the individuals are not known in advance, an unsupervised agglomerative hierarchical clustering method is considered, following the *clustering approach* of ***VIBEA*** (see Section 6.4.1). To identify groups of similar feature values clustering procedures use distance measures to group data points in a way that provides minimal inner-cluster distances and maximal inter-cluster distances [161]. Agglomerative hierarchical clustering is considered a bottom-up clustering method and follows a binary tree structure, since it starts with every single sample in a single cluster, then, in each iteration, it merges the closest pair of clusters by satisfying the aforementioned similarity criteria, until all of the data is in one cluster [209]. This technique is able to produce an ordering of the samples, which may be informative for data display, as well as the fact that smaller clusters are generated, which can be helpful for knowledge discovery [209]. Such advantages match the requirements of our problem, since the levels of mastery follow a natural order of learning through the experience acquired by the PSL speakers.

In order to gain the insights about the structure regarding the levels of mastery among the individuals, the questionnaires that all the individuals filled (see Section 6.3.3) is used. Three scales of expertise in PSL are defined, with different levels, each one based on the score of each individual's answer to a specific question, such as: i) number of years that was familiarized with sign language, (*year range*), with three levels; ii) the current profession, (*profession*), with four levels. iii) the combined score of the two answers, (*year range-profession*), a weighted combination of the scores of each answer (0.5 each), which originates four different levels. Table 6.9 shows the organization of the individuals in levels regarding their answers to the two mentioned questions.

Level	Year Range	Individuals	Level	Profession	Individuals
1	7-15	D1 H2 H7	1	Non Related to PSL	D4
2	16-25	H6 D5 D6	2	Speech Therapist	H7
3	26-35	D4	3	PSL Interpretation	H2
-	-	-	4	PSL Teaching	D1 D5 D6 H6

Table 6.9: Division of the population regarding the number of years in contact with the PSL and the current job for classification of the levels of mastery in PSL.

The clustering process aims to discover the inner structure of the data and groups the individuals in relevant clusters. To measure the clustering quality, two different metrics, that look at the class labels of the documents assigned to each cluster, are used. The first metric is the widely used *entropy* measure, that looks to how the various classes of documents are distributed within each cluster, and the second measure is the *purity*, that measures the extend to which each cluster contained documents from primarily one class.

Given a particular cluster S_r of size n_r , its *entropy* is defined by

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{6.4}$$

where q consists in the number of classes in the dataset, and n_r^i is the number of samples of the ith class that are assigned to the rth cluster. To obtain the *entropy* of the entire clustering, a sum of the individual cluster entropies weighted according to the cluster size is more informative

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \tag{6.5}$$

Ideally a clustering solution will have an *entropy* of zero. This would reveal that each cluster would contain only samples from a single class. In a similar fashion, the *purity* of the same cluster is computed as follows

$$P(S_r) = \frac{1}{n_r}max(n_r^i) \tag{6.6}$$

which is the ratio between the overall cluster size that the largest class of documents assigned to that cluster represents. The overall *purity* is given by

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r) \tag{6.7}$$

The two measures are complementary and give an indication of the clustering quality, for instance small entropy values and large purity values suggest a good clustering solution.

6.5 Validation

As stated in Section 6.4.2, the final feature vector that represents each *mini-clip* is formed by the concatenation of all the motion-based features extracted within the pre-defined ROI.

In order to obtain the MHI, a motion mask is calculated by subtracting consecutive video frames enabling the extraction of moving elements from static irrelevant background. The resulting image is then binarized, by the Otsu method [261], to obtain the final motion mask. The duration parameter of the MHI is empirically inferred and set to 10 frames for all the *mini-clips*. Fig. 6.9 shows a sequence of MHI images. The fact that the individuals are considerably distant from the camera, diminishes the quality of pixel-based approaches since the resolution is not the desired to extract the maximum detail from the frames. The presence of several isolated non-zero pixels actually confirms that fact. On the other hand, the conceptual idea of the MHI is captured, which is clearly visible by the history of the arms' motion kept on the pixel intensity.

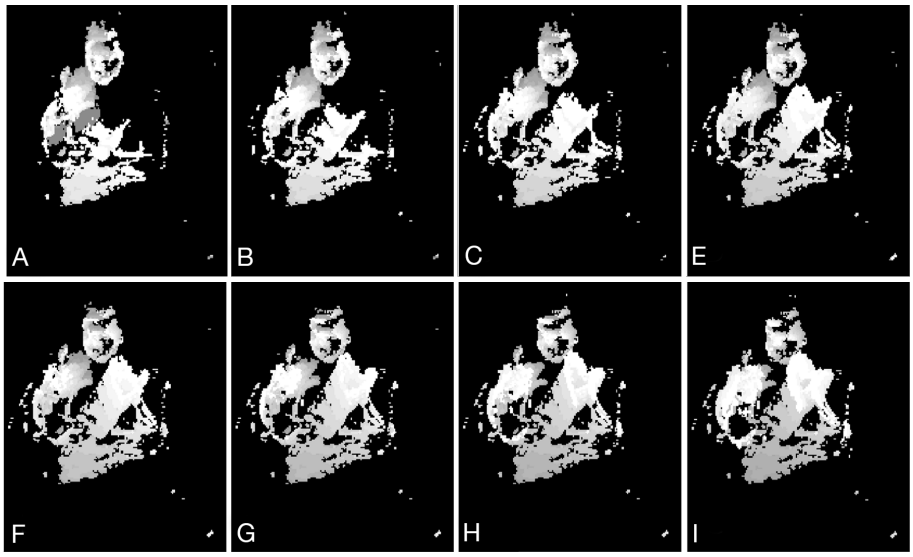


Figure 6.9: Sequence of MHI images obtained from a sequence of frames of a *mini-clip*. Higher values of pixel intensity (brighter) represent pixels in which motion occurrs more recently. The letters indicates the order for the sequence of consecutive frames.

The MG information uses a kernel of size 3×3, and the histograms of the motiongrams undergo a standardization process to allow all feature vectors for each *mini-clip* to have the same

dimension. Since the histograms of the motiongrams are features whose size, in terms of number of bins, are directly related to the resolution of the image, in this case the number of rows and columns of the ROI, the minimum values for width and height from all the *mini-clips* are considered for the normalization of the bin widths, obtaining a 114 and 162 dimensional feature vector for both vertical and horizontal motiongrams, respectively. See Fig. 6.10 for an example of vertical and horizontal motiongrams.

(a) (b)

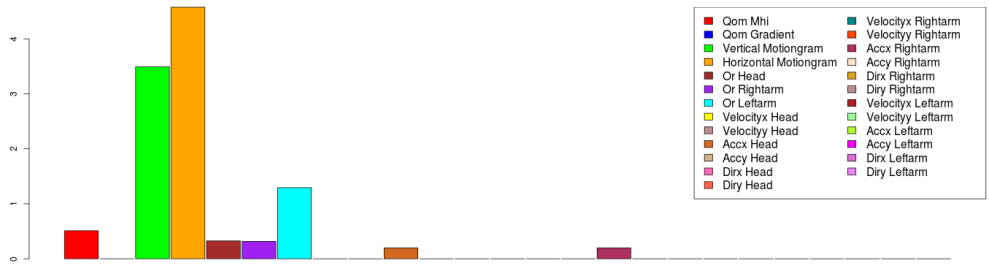
Figure 6.10: Examples of motiongrams of a *mini-clip* before normalization of bin width. (a) horizontal motiongram; (b) vertical motiongram.

Summing up all features, a 324 dimensional feature vector is obtained to represent each *mini-clip* (QoM-MHI (2) + QoM-MG (4) + Motiongram (114+162) + (Or (2) + Vel (4) + Acc (4) + Dir (4)) $\times$ 3). The multidimensional feature vector undergoes a feature selection process to inspect the redundant and inconsistent features that may affect the classification. Fig. 6.11 shows the results of the feature evaluation using the *Information Gain* [210] and *Relief-f* [166] methods. In terms of revealing the features that are most discriminative, both methods show the superior importance of the motiongrams as the ones in which more significant differences are observed between the *mini-clips*, in other words, the variance observed among all the samples is higher.

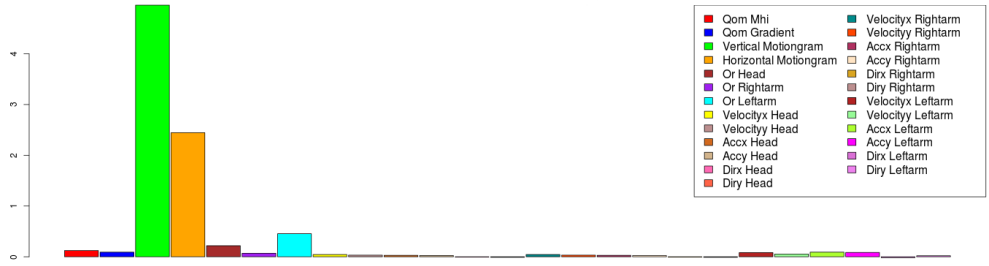
6.5.1 Distinguishing Deaf from Hearing People

In order to attempt on finding differences among the *mini-clips* of deaf and hearing people two different classifiers are used: *k-nn* and SVM. The measures used for evaluation are: a) Precision (P), b) Recall (R), c) Correct Rate (CR), The CM, which is subdivided into the *Worst CM* of each *mini-clip's* group, and the *Accumulated CM*, which is the sum of the CM that results from the average of the cross-validation process, is used to support the conclusions. For this experiment, the first row correspond to the deaf class and the second one to the hearing class.

The number of neighbors used for the *k-nn* classifier is heuristically found within the set $[3,5,7,9]$. The performance of the *k-nn* algorithm is frequently affected by problems related to the elevated number of dimensions of the considered data. Recent research indicates that the



(a)



(b)

Figure 6.11: Feature selection using the following methods: (a) *Information Gain*; and (b) *Relief-f*. In both cases, vertical and horizontal motiongrams display the highest weight.

number of dimensions alone does not necessarily result in problems finding the nearest neighbors [131], since relevant additional dimensions can also increase the contrast. The difficulties only arise when irrelevant dimensions reduce the contrast on the features. To test if irrelevant dimensions are present and also to reduce the computational cost in terms of processing time, the Principal Component Analysis (PCA) method [337] is applied to the feature vector. This way, we inspect if applying dimensionality reduction the signal-to-noise ratio is high enough to take that transformation used by PCA to represent the data. The percentage of retained variance parameter of PCA is set to 90.0%. The results of the *k-nn* with the different data groups for this experiment, *mini-clips* from each topic separately and all topics together, are shown in Table 6.10, after PCA, and Table 6.11, before PCA.

Comparing the results of both Table 6.10 and Table 6.11, mainly the CR, P and R, it is clear that the performance of the *k-nn* classifier is better when the PCA is not applied. To avoid this loss of information different normalizations such as Z-Score or Min-Max for example could be applied before using PCA [200]. However, for further experiments PCA is not applied.

Observing the *k-nn* results, the performance for all the different topics is highly satisfactory, being the highest values verified for the Topic 1. The groups are organized by *conversational topics*, not only to compare the performance of the classifier for a different number and combination of samples, but also to measure the discriminative level of the features on different emotional con-

Topic		Worst CM		Accumulated CM		CR (%)	R (%)	P (%)
T1	D	0.75	0.25	0.82	0.18	75.7	79.9	81.6
	H	0.42	0.58	0.34	0.66			
T2	D	0.69	0.31	0.75	0.25	69.1	73.7	75.3
	H	0.50	0.50	0.40	0.60			
T3	D	0.77	0.23	0.82	0.18	78.5	84.9	82.5
	H	0.36	0.64	0.29	0.71			
T4	D	0.75	0.25	0.80	0.20	73.6	82.9	79.9
	H	0.54	0.46	0.43	0.57			
All	D	0.71	0.29	0.74	0.26	66.1	74.4	73.7
	H	0.53	0.47	0.49	0.51			

Table 6.10: *k-nn* classification results obtained for the distinction between deaf and hearing people after PCA.

Topic		Worst CM		Accumulated CM		CR (%)	R (%)	P (%)
T1	D	0.97	0.03	0.97	0.03	98.9	98.9	99.4
	H	0.04	0.96	0.05	0.95			
T2	D	0.93	0.07	0.95	0.05	95.4	97.7	94.9
	H	0.05	0.95	0.04	0.96			
T3	D	0.97	0.03	0.98	0.02	97.2	97.6	98.1
	H	0.05	0.95	0.04	0.96			
T4	D	0.96	0.04	0.98	0.02	97.5	98.5	97.9
	H	0.04	0.96	0.03	0.97			
All	D	0.96	0.04	0.97	0.03	96.3	97.4	96.9
	H	0.05	0.95	0.05	0.95			

Table 6.11: *k-nn* classification results obtained for the distinction between deaf and hearing people before PCA.

texts. Topic 1 is by indication of Table 6.7 the one with less samples. The CMs show concordant results among all the groups being the misclassification rate never higher than 5.0%. No major differences are observed between the worst CM and the accumulated one which confirms the good performance values.

The same experiment is performed using binary SVM to identify the two classes: deaf and hearing. The intention is to evaluate the differences in terms of performance with a classifier less sensitive to dimensionality problems. A linear kernel function is selected from empirical experiments. Several C-SVM values are tested as input, ranging from 2^{-2} to 2^{8}, and chosen the one that optimize the performance values CR, R and P.

We can verify that the results presented in Table 6.12 corroborate the ones obtained with *k-nn* being the performance slightly improved, specially the CR measure. For the SVM, contrarily to *k-nn*, the best performance is observed for Topic 3. Regarding the CMs, it is observed a balance in the misclassification rates of the classes, which leads to conclude that the classifier is accurate for both classes. The Accumulated CM are more a less equivalent for all the topics, being the highest misclassification rates observed for the group that aggregates all the topics, where more samples

can confuse the classifier.

Topic		Worst CM		Accumulated CM		CR (%)	R (%)	P (%)
T1	D	0.95	0.05	0.97	0.03	98.8	98.7	97.9
	H	0.00	1.00	0.01	0.99			
T2	D	0.94	0.06	0.98	0.02	98.3	98.3	98.2
	H	0.00	1.00	0.00	1.00			
T3	D	0.97	0.03	0.99	0.01	98.6	99.2	99.0
	H	0.06	0.94	0.01	0.99			
T4	D	1.00	0.00	1.00	0.00	98.6	98.5	99.0
	H	0.06	0.94	0.03	0.97			
All	D	0.94	0.06	0.98	0.02	97.5	96.7	96.1
	H	0.18	0.82	0.04	0.96			

Table 6.12: SVM classification results obtained for the distinction between deaf and hearing people.

6.5.2 Distinguishing Different Conversational Topics

Since the individuals are speaking PSL in all the *conversational topics*, only a PSL speaker could evaluate the contents of the videos and verify if the four distinct moments are in fact present. Therefore, this supervision was pursued by two PSL experts, and they confirmed that the individuals were demonstrating different levels of expressiveness and emotion according to the current discussion topic. To find out if it is possible with the selected features to discriminate the expressiveness of the individuals on the *mini-clips* of those moments, the same classification procedure from the previous experiment is conducted, as well as the same parameters tuning for the classifiers.

Table 6.13 shows that the performance of the *k-nn* classifier is pretty good, since the CR for all the groups is around 90.0%. Even considering the group that aggregates all the individuals, the performance keeps a high classification rate, CR= 92.9%, R= 91.1% and P= 93.5%. Such results confirm that the selected features permits the distinction between the four *conversational topics*, enabling the identification of different motion expressions. Inspecting the CMs, it is possible to observe a pattern, which is that Topics 1 and 2 display the highest misclassification rates. This is not a strange occurrence if we look at what the individuals are supposed to talk about on both topics. Indeed, Topics 1 and 2 are very similar in terms of content and emotional context. This may be an explanation for the observed occurrence.

Table 6.14 shows that performance values of the multiclass SVM classifier are somewhat different from the ones of the *k-nn*. We reminded that in this experiment a multi-class SVM for ordinal data [289] is used. While on the *k-nn* the performance results are fairly homogeneous among the different groups, this is not observed for the SVM. Considering the CM, a particular phenomenon is observed, which is the fact that the misclassifications happen mainly in adjacent classes, which is a consequence of the fact that the classes are considered in an ordered way, as mentioned in Section 6.4.4. Due to this fact, the classifier in case of doubt decides for the label to

Individual		Worst CM				Accumulated CM				CR (%)	R (%)	P (%)
D1	T1	0.77	0.06	0.03	0.13	0.86	0.06	0.01	0.07	91.4	86.0	90.7
	T2	0.03	0.90	0.03	0.05	0.02	0.92	0.02	0.04			
	T3	0.02	0.00	0.95	0.03	0.01	0.02	0.94	0.03			
	T4	0.10	0.07	0.00	0.83	0.04	0.03	0.01	0.92			
D4	T1	0.94	0.06	0.00	0.00	0.87	0.08	0.06	0.00	96.4	94.8	91.8
	T2	0.33	0.67	0.00	0.00	0.02	0.88	0.02	0.08			
	T3	0.00	0.00	1.00	0.00	0.02	0.00	0.97	0.01			
	T4	0.00	0.00	0.00	1.00	0.00	0.02	0.00	0.97			
D5	T1	0.81	0.10	0.07	0.02	0.87	0.08	0.06	0.00	93.3	86.6	93.1
	T2	0.03	0.90	0.01	0.06	0.02	0.88	0.02	0.08			
	T3	0.00	0.00	0.99	0.01	0.02	0.00	0.97	0.01			
	T4	0.00	0.03	0.00	0.97	0.00	0.02	0.00	0.97			
D6	T1	0.89	0.01	0.04	0.06	0.95	0.01	0.03	0.01	93.6	95.2	96.4
	T2	0.00	0.81	0.08	0.11	0.03	0.87	0.06	0.05			
	T3	0.00	0.00	0.97	0.03	0.00	0.01	0.97	0.02			
	T4	0.04	0.04	0.01	0.92	0.01	0.03	0.03	0.92			
H2	T1	0.81	0.11	0.06	0.03	0.93	0.04	0.02	0.00	91.7	93.6	98.3
	T2	0.00	1.00	0.00	0.00	0.02	0.94	0.03	0.00			
	T3	0.00	0.00	0.96	0.04	0.00	0.05	0.91	0.04			
	T4	0.00	0.04	0.04	0.92	0.00	0.05	0.06	0.89			
H6	T1	0.72	0.25	0.03	0.00	0.84	0.15	0.01	0.00	89.8	83.9	89.6
	T2	0.11	0.80	0.09	0.00	0.07	0.86	0.07	0.01			
	T3	0.02	0.06	0.92	0.00	0.01	0.05	0.92	0.01			
	T4	0.00	0.00	0.03	0.97	0.00	0.01	0.03	0.96			
H7	T1	0.84	0.13	0.00	0.03	0.94	0.04	0.00	0.01	97.0	94.6	99.9
	T2	0.00	0.96	0.02	0.02	0.00	0.95	0.03	0.02			
	T3	0.00	0.04	0.96	0.00	0.00	0.01	0.99	0.00			
	T4	0.00	0.04	0.00	0.96	0.00	0.01	0.00	0.98			
All	T1	0.90	0.03	0.04	0.03	0.91	0.03	0.03	0.03	92.9	91.1	93.5
	T2	0.03	0.91	0.02	0.04	0.03	0.91	0.02	0.04			
	T3	0.01	0.01	0.96	0.03	0.01	0.01	0.96	0.03			
	T4	0.01	0.03	0.03	0.93	0.01	0.04	0.03	0.92			

Table 6.13: *k-nn* classification results obtained for the distinction the different *conversational topics*.

assign between the adjacent classes. Therefore, and contrarily to the *k-nn* classifier, Topics 1 and 2 do not display the highest rates of misclassification. However, both classifiers present high CR, P and R values, which reveals the descriptive and discriminative potential of the extracted motion features set.

6.5.3 Identifying Levels of Mastery in PSL

As stated in Section 6.4.5, the solution to identify different levels of mastery in PSL pass by an agglomerative hierarchical clustering algorithm that can be able to reveal the data structure and measure the quality of the clustering process through the quality of each cluster and the degree of similarity between elements in the same cluster. Details about the algorithm used can be found in [161].

Individual		Worst CM				Accumulated CM				CR (%)	R (%)	P (%)
D1	T1	0.80	0.20	0.00	0.00	0.83	0.17	0.00	0.00	91.7	92.2	90.2
	T2	0.08	0.92	0.00	0.00	0.08	0.93	0.00	0.00			
	T3	0.00	0.05	0.95	0.00	0.00	0.02	0.98	0.00			
	T4	0.00	0.00	0.10	0.90	0.00	0.00	0.13	0.87			
D4	T1	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	98.6	98.9	97.9
	T2	0.17	0.83	0.00	0.00	0.09	0.91	0.00	0.00			
	T3	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00			
	T4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00			
D5	T1	0.88	0.13	0.00	0.00	0.82	0.18	0.00	0.00	91.0	90.4	89.6
	T2	0.15	0.85	0.00	0.00	0.08	0.88	0.04	0.00			
	T3	0.00	0.08	0.85	0.08	0.00	0.05	0.92	0.03			
	T4	0.00	0.00	0.05	0.95	0.00	0.00	0.04	0.96			
D6	T1	0.85	0.15	0.00	0.00	0.92	0.08	0.00	0.00	90.6	89.6	90.0
	T2	0.00	0.71	0.29	0.00	0.03	0.86	0.11	0.00			
	T3	0.00	0.13	0.81	0.06	0.00	0.06	0.88	0.05			
	T4	0.00	0.00	0.12	0.88	0.00	0.00	0.07	0.93			
H2	T1	0.93	0.07	0.00	0.00	0.97	0.03	0.00	0.00	86.6	86.8	85.3
	T2	0.22	0.67	0.11	0.00	0.11	0.74	0.16	0.00			
	T3	0.00	0.15	0.85	0.00	0.00	0.12	0.88	0.00			
	T4	0.00	0.00	0.17	0.83	0.00	0.00	0.17	0.83			
H6	T1	0.90	0.10	0.00	0.00	0.97	0.03	0.00	0.00	98.1	98.5	97.7
	T2	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00			
	T3	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00			
	T4	0.00	0.00	0.09	0.91	0.00	0.00	0.06	0.94			
H7	T1	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	97.9	98.4	98.2
	T2	0.00	0.83	0.17	0.00	0.00	0.93	0.07	0.00			
	T3	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00			
	T4	0.00	0.00	0.05	0.95	0.00	0.00	0.02	0.98			
All	T1	0.50	0.50	0.00	0.00	0.70	0.29	0.01	0.00	78.8	82.1	77.9
	T2	0.08	0.58	0.33	0.00	0.04	0.76	0.19	0.01			
	T3	0.00	0.14	0.86	0.00	0.01	0.08	0.86	0.06			
	T4	0.00	0.00	0.40	0.60	0.00	0.00	0.20	0.80			

Table 6.14: SVM classification results obtained for the distinction the different *conversational topics*.

For this experiment, there are three scales that define the criterion from which the number of clusters are set: i) group by *year range*, three clusters; ii) group by *profession*, four clusters; iii) group by *year range-profession*, four clusters. Table 6.15, Table 6.16, and Table 6.17 show the statistics that evaluate the clustering performance for each scale, respectively. The column labeled as C_{id} indicates the id of the cluster, *Size* the number of samples that belong to each cluster, *ISim* and *ISdev* represent both the average and standard deviation in terms of similarity between each cluster (internal similarities), respectively, whereas *ESim* and *ESdev* represent the same statistics but for similarity of the samples of each cluster and the rest of the samples (external similarities), respectively. It is important to note that small values of C_{id} indicate that clusters are tight and far away from the rest of the samples. For this experiment, although not being a classification problem, we also use the CM, from association of cluster labels with ground-truth,

as a supplementary source to support conclusions.

Regarding the *year range* scale (see Table 6.15), the PSL experts report that such indicator alone may be an ambiguous criterion to define how experienced an individual is in terms of PSL knowledge. This happens since many factors, regardless of the years in contact with the language, may influence this: for example, if an individual has to use sign language on a daily basis or not, if he has parents who are deaf making more likely that language to be the one that the individual considers to be his mother tongue, whether he was born hearing or deaf, and in the case of being deaf if the deafness was acquired or from birth, among other facts. All the mentioned factors show that the years in contact with the language is not an unequivocal indicator since the life experience of the individual may overcome that issue. The overall entropy and purity of this clustering are 0.66 and 0.64, respectively. These values reveal that the clustering is performed poorly since the entropy is around 0.7. Purity which is supposed to be as close as 1.0 as possible is quite low which indicates that the samples in each cluster are not as homogeneous as desired. Analysing the CM, we verify that cluster 3, although it is the one with more elements, it is the cluster with less samples assigned to it and further away from the rest of the samples. Clusters 1 and 2 also display a large percentage of misclassification.

C_{id}	Size	ISim	ISdev	ESim	ESdev	Entropy	Purity	CM		
1	70	0.67	0.08	0.33	0.08	0.63	0.56	0.44	0.56	0.00
2	167	0.61	0.11	0.48	0.12	0.60	0.71	0.71	0.28	0.01
3	1029	0.60	0.11	0.44	0.12	0.75	0.65	0.28	0.65	0.07

Table 6.15: Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes are regarding the number of years in contact with PSL (*year range*).

When performing the clustering by *profession* scale (see Table 6.16), the overall entropy and purity of the solution are the best of the three analysis performed (0.59 and 0.68, respectively). From the questions featured on the questionnaires that are common to both deaf and hearing, the one regarding the current professional occupation of the individuals is the one considered by PSL experts to possibly be the most discriminative when it comes to the expertise on this language. This is corroborated by an improvement on the values of entropy and purity obtained for this scale. The CM reveals that the overall misclassification is lower, being observed a higher rate of well classified samples comparing to the misclassified ones of the first cluster.

C_{id}	Size	ISim	ISdev	ESim	ESdev	Entropy	Purity	CM			
1	70	0.67	0.08	0.33	0.08	0.59	0.60	0.60	0.36	0.00	0.04
2	167	0.61	0.11	0.48	0.12	0.53	0.70	0.69	0.28	0.01	0.02
3	144	0.71	0.08	0.56	0.09	0.70	0.67	0.67	0.12	0.18	0.03
4	885	0.60	0.11	0.48	0.13	0.54	0.74	0.74	0.01	0.05	0.20

Table 6.16: Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes are regarding the current profession of the individual (*profession*).

The clustering by *year range-profession* scale (see Table 6.17) intents to obtain the clusters that concatenate more information about the expertise of each subject. However, such fusion reveals the less promising results of entropy and purity (0.76 and 0.52, respectively), which indicates that this combination impairs the performance of the clustering solution. The CM shows that the samples are mainly distributed for the first three clusters. The highest value of entropy and the lowest value of purity are confirmed by the observation of a large number of misclassifications for all clusters, especially for the first one. By aggregating more information, it is expected a more detailed description of individuals, however, this is not observed. To overcome this issue a different fusion of the information should be tested.

C_{id}	Size	ISim	ISdev	ESim	ESdev	Entropy	Purity	CM			
1	70	0.67	0.08	0.33	0.08	0.70	0.56	0.04	0.56	0.36	0.04
2	167	0.61	0.11	0.48	0.12	0.83	0.41	0.41	0.28	0.29	0.02
3	144	0.71	0.08	0.56	0.09	0.85	0.42	0.25	0.42	0.30	0.03
4	885	0.60	0.11	0.48	0.13	0.66	0.68	0.06	0.68	0.06	0.20

Table 6.17: Clustering statistics and performance measures obtained by the agglomerative clustering solution. The classes are the number of years in contact with PSL combined with the current profession (*year range-profession*).

6.6 Summary

Deaf people who are sign language speakers are individuals considered to be extremely expressive. The lack of ability to use spoken language causes these individuals to rely much more on other means to convey their expressive intentions and emotions, which may cause social inclusion in specific contexts. Automated visual analysis of behavior provides tools for the construction of intelligent computer vision systems. Digital applications supported by analysis of expressiveness in social context may be a solution for deaf people to express and integrate into society in a better way.

This Chapter presents a body expressiveness analysis framework, ***VIBEA***, to address the research questions exposed in Section 6.1, which are defined taking in consideration the main social needs of the PSL population, and the core technical concepts to achieve a sensitive behavioral and context-based model. A novel dataset that deals with a very specific context, duo-interaction between deaf and hearing people, is presented and supported by sociological and psychological principles and supervision. It is being prepare to be made publicly available, since, despite its specificity, it can also be valuable for other studies among the scientific community.

Despite the big effort to construct this dataset and its novelty, the biggest drawback in this study is the size of the sample set, which may not offer enough evidence for the obtained results. A larger sample set is required to sustain ***VIBEA***. However, important conclusion may be outlined regarding the research questions. It is possible to distinguish with high recognition rates the deaf from hearing people among the individual in the sample set, as well as the differentiation between

several *conversational topics*. Such outcomes indicate that the body features extracted, and their combination, are descriptive and discriminative enough to be tested in further experiments related with human expressiveness. These conclusion gains more credibility if we consider that the consulted PSL experts stated that facial features are the main cues to distinguish a deaf from a hearing person, and that deaf people is more expressive facially than hearing people. However, space is left to explore facial attributes, their fusion, and more complex classification schemes.

Concerning the distinction among levels of mastery in PSL, the current profession is the question featured on the questionnaires that most suitably describes the individual's expertise in PSL. However, more information is needed on how to obtain evidence to represent the levels of mastery in PSL from multimodal data. Indeed, identifying the level of mastery of each individual is not linear (even for a human this process can be ambiguous and erroneous), being dependent on a numerous amount of factors that may not be considered by the questions featured in the questionnaires. Further work on this matter is necessary to identify the more relevant indicators of the expertise of a person in PSL, such as a multi-variable regression for which we would need to define other questionnaires and tests with combinations of demographical, academical and other data that could evaluate more accurately the expertise of an individual.

Summing up, the main breakthrough of this study is that, contrarily to what was expected, body gestures alone contain a great amount of expressive content that allowed us to identify valuable aspects in the videos made available by the construction of the novel dataset. ***VIBEA*** represents our last contribution in this thesis, closing the loop of the study of motion and context representations in several settings at different levels of human activity.

Chapter 7

Conclusions

Valid criticism does you a favor.

Carl Sagan

Inferring human activity automatically from video is a dense and complex problem that spans from low-level feature detection to high-level semantics interpretation. Starting with nothing more than simple pixels, the literature has shown how hierarchically connected layers of processing can produce results with application on such areas as monitoring and detection of motion patterns, understanding social behavior, multimedia video classification or body expressiveness analysis. But by nature of the raw material, and a research bias towards specific objectives, a reliable solution for all types of applications involving human activity inference is still not a reality. By grouping human movement in levels, and presenting solutions at each level, this thesis aims for a step further on the solution.

7.1 Discussion

This thesis has presented an investigation of human activity in computer vision at different levels. Our work addresses the challenges of different domain settings to propose four frameworks based on motion as main driving force and context as supportive source. Each domain settings provides specific conditions, both technical-based and content-based, that enable the proper context to study relevant topics, such as tracking, person detection, among others, of human activity at different perceptual levels, *the group*, *the whole* and *the parts*. Such levels motivate our work and permit us to find the appropriate settings and, consequently, the individual research gaps to tackle in each one. Our methodology follows an application-based line, where real-world problems are identified and an evaluation has been carried out.

In this Chapter, we divide our discussion by the aforementioned human activity levels. On each is presented the respective frameworks along with an overall evaluation about their contributions, limitations and improvements. A more generic approach is presented on the potential directions for future work.

7.1.1 The Group

The level *group* in our thesis has a dual-meaning: global motion patterns and social collective activity. The increasing demanding for both topics in the literature is undeniably, and surveillance settings have been identified as the preferred scenario of analysis [2], since it provides a wide view of the scene, permits to analyze the environment with several pedestrians and object of interest.

A first challenge is to provide a solution *to the lack of a generic long-range motion representation* with the aim to *enhance the analysis of spatio-temporal patterns and extract meaningful motion statistics* in different surveillance scenarios. We proposed a motion analysis framework, ***VILOMA***, that embeds local and global motion information to capture longer spatio-temporal changes in the scene. It was develop in order to: collect, quantize and refine information into local cells, and combine their neighborhood data using information theory principles; advect global motion properties through temporal integration of dense flow. ***VILOMA*** permits to extract long-range global trajectories from segments of the video sampled uniformly in time. It includes a new global outlier removal technique for flow vector data to improve the refinement of local motion information. It also considers a re-correlation algorithm at the tracklet-level, with global optimization, to link broken trajectories.

Its major pertinence is its generic approach for different scenarios, whose content, in terms of number of people and randomness of movements, differ substantially among them. However, its computational effort is demanding due to the temporal integration of flow. For surveillance applications that allow offline processing, this is not a relevant issue, but it will be interesting to parallelized the integration (since the particles are independent components) in order to achieve real-time or to implement a streaming mode of the framework. The global trajectories extracted by ***VILOMA*** at the end of each segment are indeed close to the individual trajectories of pedestrians, but their main advantage are their spatial range and density. It was shown in Chapter 3 the usefulness of such representation for motion segmentation, where the results largely surpass the state-of-the-art. However, they lack temporal information. At this moment, it is possible to know the starting and ending frames, but it is not possible to infer the in-between information temporally. Probably, again, a streaming mode of the framework could attain more temporal information. ***VILOMA*** along its pipeline builds rich motion information in terms of physical dynamics, such as streak flows, and representation, such as the fine-to-coarse flow, that were not explored yet. We believe that this information can provide advantages for human activity-task related. As last comment, more tests need to be done not only in more datasets, but mainly to infer the impact of individual modules in the final behavior of the framework.

The second challenge is to provide a *relational representation of social dynamics* with the aim to *classify individual and collective behavior in an environment* for surveillance awareness. First, new semantic concepts that lack in the technical literature and that are supported by social-psychology principles were proposed to the computer vision literature. As well, a novel surveillance dataset was extended with such semantic labels, along with low-level information regarding the adopted features. Then, we proposed a social behavior identification framework, ***VISOBI***,

that identifies meaningful relational features among the entities of interest, namely individuals and objects, and creates a descriptor that encodes, at multiple resolutions, their relationships. ***VISOBI*** enlarges its potentiality not only for the classification of individual and collective behavior, but also for the relevance analysis of the individual features in a two-fold manner: impact-factor on final classification, and sociological meaning. An exhaustive evaluation was undertaken highlighting the benefits and exposing the drawbacks of the framework.

One of the main advantages of ***VISOBI*** is its mini-batch approach to represent and encode the relational information acquired from the scene. This dynamic behavior is the first step for an online classification framework. However, we verified a high impact of the low-level processing modules, especially tracking, in the final classification performance. Such modules should be revised internally or extended with external information. One of the main conclusions that we can draw is that the annotation process should be reviewed. Despite all the effort involved in defining the concepts, stating the technical-socio-psychological principles, adopting confidence measures and rules to validate the annotation, some errors were verified and some obtained results point out annotation ambiguities. Some of the inconsistency comes from usual technical problems, such as resolution and size of the image, but we believe that most of it comes from the subjectiveness of select the correct I.P. or G.B. label. Therefore, a new approach based on soft labelling assignment with different annotations from various users could be an interesting line of work. Following this line of work, the annotation should be extended to the other videos of our dataset, and other datasets more common in the literature, in order to improve the validation of the semantic concepts and test our framework in other conditions. A larger collaborative effort with experts in the areas of sociology and psychology should be considered for a complete validation of ***VISOBI*** in their areas.

Both frameworks lies on the same scenario, but deal with different questions. However, their structure and outcomes may be complementary. This connection was not done in this thesis, since it would probably imply to dispense the study on the remaining *levels* of human activity. However, we outline here some lines of work for their integration. For instance, ***VILOMA*** may extract structural information from the scene through the characteristics of the extracted trajectories, such as regions of interest, common paths, entry/exit regions, areas of (un)structured movement, etc. This information may be reused by ***VISOBI*** to automatically update the relations between individuals and objects of interest, as well as to integrate new information, e.g. relation between individuals exit-regions, to enhance its relational representation. The flow characteristics of ***VILOMA*** may also be integrated into *VISOBI*, for instance in the motion model of the tracking module to improve its performance. An opposite feedback loop can be also considered. For instance, the individual information of pedestrians in *VISOBI* may be used by the local cells, in ***VILOMA***, to refine the representative flow vectors and, consequently, improve the generation of the global motion trajectories. As final remark, the integration of both frameworks is pertinent and may lead to a robust representation of the *group* level.

7.1.2 The Whole

The level *whole* in our thesis represents the combination of kinematics-foreground with contextual-background descriptors to represent the action and the scene in the frame. The exponential growth of the user generated online videos have been leading to an increased interest in the research community to provide solutions for the automatic classification of multimedia videos [335].

The third challenge is to provide a solution *to identify relevant movement* with the aim *to correctly integrate contextual information and increase the classification performance* in multimedia videos. We proposed an action recognition framework, ***VIMUAR***, that combines known concepts embedded into novel formulations. Three major concepts are investigated and included as modules in the pipeline of ***VIMUAR***. First, the camera modelling and motion compensation module benefits from a complementary representation of the camera, extracting both its global parametric dynamics model and local statistics from a grid of cells in the image plane. One advantage is the possibility to detect camera changes in terms of categories pre-defined, which means a significant change of camera context. We have used this *trigger* as frequency to sample the contextual descriptors, in this case, a descriptor derived by the characteristics of this module. Second, a motion-based saliency map that separate relevant regions, foreground, from non-relevant regions, background. The main advantage stated in our experiments was the reduction of the number of trajectories, without impairing the performance. Further quantitative and qualitative experiments should be done to evaluate its real contribution from the motion camera compensation module, since both were tested together. Our intuition is that, if indeed the separation between foreground and background is effective, the inclusion of image descriptors that characterize the scene, sampled in the background regions, should provide a significant improvement in the classification. Finally, the video-shot segmentation and summarization module aims mostly to provide an effective mean to drastically reduce the computational effort while keeping fair classification results. Again, further experiments need to be taken to infer the correct behavior of this module. The recent outcomes show promising and very satisfactory results, indicating this approach as a good alternative to the most known *keyframe-based* and *video-based* approaches.

Multimedia video classification is a high demanding problem, not just in terms of execution, since the datasets are very large and complex, but also in terms of system functionality. Therefore, it is very important to keep the modules isolated with well-defined inputs and outputs. Apart from this technical comment and further experiments that should be performed, we must investigated the inclusion of spatio-temporal segmentation structures in ***VIMUAR***. Some of the existing methods benefit from longer temporal segments as inputs, since their formulation permits to improve the long-term temporal coherence in their regions and boundaries [114]. In this way, a shot correctly segmented with coherent content may be a good candidate for the spatio-temporal segmentation algorithm. Exploring spatial and temporal correlations within the descriptors that represent the trajectories volume can also be a good alternative to increase the classification accuracy.

7.1.3 The Parts

The level *parts* in our thesis represents articulated regions of the body that express intention and meaning in non-verbal communication. Its pertinence comes from the increasing interest for computer vision and artificial intelligence researchers on developing computer systems and models to achieve nonverbal sensitivity in different behavioral settings and through different channels such as body gesture [60].

The fourth challenge is to provide a *combination of motion features* with the aim to *capture expressiveness intention and characterization* for non-verbal behavioral settings. The presented study is conducted over an overall project that aims to offer a digital mean to mitigate social inclusion of deaf people. Our proposal works on one branch of that project covering a framework for body expressiveness analysis, ***VIBEA***, of a duo-interaction between deaf and hearing people in several emotional contexts, called *conversational topics*. A novel dataset was created with advise of experts in the fields of PSL (Portuguese Sign Language) and socio-psychology. The venue, physical recording setup, sample set characteristics, questionnaires, scenarios, and conversational topics were carefully prepared to provide the necessary conditions to address the designed needs, namely differentiation of deaf and hearing people, identification of different conversational topics, and identification of different levels of mastery of PSL speakers.

The major drawback of this study is the number of samples in the dataset. The obtained results are highly positive, however the conclusions drawn need to be validated with more samples. Therefore, is mandatory to extend the dataset. However, the efforts are tremendous, not only for the technical resources involved, but also, and mainly, in finding volunteers and a common venue to ensure the recordings. An alternative should be considered but also the collaboration with the experts in the areas of sociology and psychology should be kept. Despite the lack of enough samples, the high recognition rates indicate that the selected combination of kinematics features may be descriptive and discriminative enough to represent body expressiveness and its variation according with emotional context. Further tests could be done with the proposed framework in other type of datasets that present simpler and well-identified expressions in order to assess their classification performance. Despite the framework trace results higher than expected in this kind of scenario, since the facial attributes assume a crucial role stated by the literature and corroborated by the knowledge of the PSL advisors, for sure the integration of features from other channel such as voice and face may help to expand and improve the expressiveness framework for multimodal behavioral data. Finding the correct hierarchical structure of mastery levels of PSL is very abstract and a more extensive preliminary study should be done to ensure which data sources should be considered for further processing.

7.2 Future Work

Although the studies of this thesis presented important contributions and some conclusive results, and despite the multipart nature of the thesis has been dealt with a well-defined goal, the presented

work is far from being exhaustive. We foresee several research directions brought up by our work either in each defined level of human activity or in a combination of them. First the concern is to exploit our findings and explore improvements in the proposed frameworks individually. Some of them were already mentioned in the previous Section while outlying further improvements. Therefore remains tracing a line of investigation that could draw a solution for at least two of the levels studied in this thesis. Such exercise is not easy to accomplish since by our definitions the level of human action is closely connected to the domain settings. In those terms, we have verified that the differences among each scenario are clear and a shared solution may be very complicated to achieve. It worths to mention the effort to automate almost every single module of the proposed frameworks. It is not easy and straightforward to control and measure every step of a complex system, even with a module-based architecture. In this way, as general remark we have stated that visual human-activity questions are complex by nature and that their solutions may benefit from complex frameworks, but not in terms of complexity added to the modules, but in terms of wisely link simple modules that take into consideration complementary features. This is the unified and general approach that we may suggest as future research in the area of human activity analysis.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560, 2008.

[2] P. Afsar, P. Cortez, and H. Santos. Automatic visual detection of human behavior: a review from 2000 to 2014. *Expert Systems with Applications*, 2015.

[3] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011.

[4] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[5] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. In *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 2–14. IEEE, 1994.

[6] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Nonrigid motion analysis: Articulated and elastic motion. *Computer Vision and Image Understanding*, 70(2):142–156, 1998.

[7] J. K. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 640–647. IEEE, 2004.

[8] P. E. Agre. The dynamic structure of everyday life. Technical report, DTIC Document, 1988.

[9] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: Its variants and applications. *Mach. Vision Appl.*, 23(2):255–281, Mar. 2012.

[10] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. Subrahmanian, P. Turaga, and O. Udrea. A constrained probabilistic petri net framework for human activity detection in video. *Multimedia, IEEE Transactions on*, 10(6):982–996, 2008.

[11] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–6, 2007.

[12] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*. IEEE Computer Society, 2007.

[13] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *Computer Vision–ECCV 2008*, pages 1–14. Springer, 2008.

[14] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):288–303, 2010.

[15] J. F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994.

[16] A. Álvarez-Meza, D. Cárdenas-Peña, and G. Castellanos-Dominguez. Unsupervised kernel function building using maximization of information potential variability. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 335–342. Springer, 2014.

[17] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Computer Vision–ECCV 2012*, pages 187–200. Springer, 2012.

[18] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69(2):159–180, 2006.

[19] I. Ari and L. Akarun. Facial feature tracking and expression recognition for sign language. In *Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th*, pages 229–232, 2009.

[20] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[21] A. Atkinson, W. Dittrich, A. Gemmell, and A. Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33:717–746, 2004.

[22] A. P. Atkinson, M. L. Tunstall, and W. H. Dittrich. Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition*, 104(1):59 – 72, 2007.

[23] S. Ayache, G. Quénot, and J. Gensel. *Classifier fusion for SVM-based multimedia semantic indexing*. Springer, 2007.

[24] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, 2009.

[25] R. V. Babu and K. Ramakrishnan. Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision Computing*, 22(8):597 – 607, 2004.

[26] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *In Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[27] F. Bashir, A. Khokhar, D. Schonfeld, et al. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, 2007.

[28] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, 2010.

[29] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: A social signaling perspective. In *Video Analytics for Business Intelligence*, volume 409, pages 271–305. 2012.

[30] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[31] Y. Benabbas, N. Ihaddadene, and C. Djeraba. Motion pattern extraction and event detection for automatic visual surveillance. *Journal on Image and Video Processing*, 2011:7, 2011.

[32] B. Benfold and I. Reid. Colour invariant head pose classification in low resolution video. In *BMVC*, pages 1–10, 2008.

[33] A. Bera, D. Wolinski, J. Pettré, and D. Manocha. Real-time crowd tracking using parameter optimized mixture of motion models. *arXiv preprint arXiv:1409.4481*, 2014.

[34] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *J. Image Video Process.*, 2008:1:1–1:10, Jan. 2008.

[35] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[36] S. Birchfield. Klt: An implementation of the Kanade-Lucas-Tomasi feature tracker. http://www.ces.clemson.edu/stb/klt/, 2007.

[37] M. J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1. IEEE, 1999.

[38] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, Mar 2001.

[39] A. F. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1257–1265, 1997.

[40] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.

[41] K. Bousmalis, M. Mehu, and M. Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203–221, 2013.

[42] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(7):1030–1044, 1999.

[43] G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. In *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, pages 238–244, 2000.

[44] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 568–574. IEEE, 1997.

[45] M. Breitenstein, H. Grabner, and L. Van Gool. Hunting nessie - real-time abnormality detection from webcams. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1243–1250, Sept 2009.

[46] F. Brémond, M. Thonnat, and M. Zúniga. Video-understanding framework for automatic behavior recognition. *Behavior Research Methods*, 38(3):416–426, 2006.

[47] A. J. Brockmeier, J. S. Choi, E. G. Kriminger, J. T. Francis, and J. C. Principe. Neural decoding with kernel-based metric learning. *Neural computation*, 26(6):1080–1107, 2014.

[48] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 594–601. IEEE, 2006.

[49] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, *ECCV (4)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36. Springer, 2004.

[50] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 282–295, Berlin, Heidelberg, 2010. Springer-Verlag.

[51] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):500–513, Mar. 2011.

[52] J. Burgoon, M. Adkins, D. N. Metaxas, R. E. Younger, J. Kruse, M. L. Jensen, T. Meservy, D. P. Twitchell, A. Deokar, J. F. Nunamaker, et al. An approach for intent identification by building on deception detection. In *null*, page 21a. IEEE, 2005.

[53] J. K. Burgoon, M. L. Jensen, T. O. Meservy, J. Kruse, and J. Nunamaker. Augmenting human identification of emotional states in video. In *Intelligence Analysis Conference, McClean, VA*, 2005.

[54] H. Buxton. Learning and understanding dynamic scene activity: a review. *Image and vision computing*, 21(1):125–136, 2003.

[55] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. In A. Camurri and G. Volpe, editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 2915 of *Lecture Notes in Computer Science*, pages 20–39. Springer Berlin Heidelberg, 2004.

[56] G. A. Carpenter, S. Grossberg, and J. H. Reynolds. Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural networks*, 4(5):565–588, 1991.

[57] D. E. Cartwright and A. E. Zander. Group dynamics research and theory. 1953.

[58] J. Cassell. A framework for gesture generation and interpretation. In *Computer Vision in Human-Machine Interaction*, pages 191–215. Cambridge University Press, 2000.

[59] C. Cedras and M. Shah. Motion-based recognition a survey. *Image and Vision Computing*, 13(2):129–155, 1995.

[60] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[61] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Appearance-based head pose estimation with scene-specific adaptation. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1713–1720, November 2011.

[62] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[63] M.-C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 747–754. IEEE, 2011.

[64] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[65] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289. IEEE, 2009.

[66] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280. IEEE, 2011.

[67] S. Cohen. Background estimation as a labeling problem. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1034–1041. IEEE, 2005.

[68] R. T. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, et al. *A system for video surveillance and monitoring*, volume 2. Carnegie Mellon University, the Robotics Institute Pittsburg, 2000.

[69] C. Cortes, M. Mohri, and A. Rostamizadeh. L 2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2009.

[70] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. pages 1071–1076, 1995.

[71] A. Criminisi. *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. PhD thesis, University of Oxford, Dept. Engineering Science, 1999. D.Phil. thesis.

[72] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. D. Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *Proceedings of the British Machine Vision Conference*, pages 23.1–23.12. BMVA Press, 2011. http://dx.doi.org/10.5244/C.25.23.

[73] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):781–796, 2000.

[74] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[75] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer, 2006.

[76] C. Darwin. *The expression of the emotions in man and animals*. London: John Murray, 1872.

[77] G. Daza-Santacoloma, J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz, and G. Castellanos-Domínguez. Dynamic feature extraction: an application to voice pathology detection. *Intelligent Automation & Soft Computing*, 15(4):667–682, 2009.

[78] B. de Gelder and J. V. den Stock. The bodily expressive action stimulus test (beast). construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Frontiers in Psychology*, 2:181, 2011.

[79] P. De Silva, M. Osano, A. Marasinghe, and A. Madurapperuma. Towards recognizing emotion with affective dimensions through body gestures. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 269–274, 2006.

[80] P. R. De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Computer Animation and Virtual Worlds*, 15(3-4):269–276, 2004.

[81] P. R. De Silva, M. Osano, A. Marasinghe, and A. P. Madurapperuma. Towards recognizing emotion with affective dimensions through body gestures. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 269–274. IEEE, 2006.

[82] J. V. den Stock and R. Righart. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 3:487–494, 2007.

[83] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1990–1997. IEEE, 2010.

[84] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[85] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.

[86] M. Douze, D. Oneata, M. Paulin, C. Leray, N. Chesneau, D. Potapov, J. Verbeek, K. Alahari, C. Schmid, L. Lamel, et al. The inria-lim-vocr and axes submissions to trecvid 2014 multimedia event detection. 2014.

[87] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1338–1345. IEEE, 2012.

[88] P. Ekman. Are there basic emotions. *Psycological Review*, 99:550–553, 1992.

[89] P. Ekman and W. V. Friesen. *Unmasking The Face*. Prentice-Hall, 1975.

[90] P. Ekman and W. V. Friesen. *Facial action coding system:A technique for the measurement of facial movement.* Consulting Psychologists Press., 1978.

[91] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[92] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

[93] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian conference on Image analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[94] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving" bag-of-keypoints" image categorisation: Generative models and pdf-kernels. 2005.

[95] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[96] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[97] B. Fergani et al. Evaluating a new classification method using pca to human activity recognition. In *2013 International Conference on Computer Medical Applications (ICCMA)*, 2013.

[98] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[99] D. Forsyth. *Group dynamics*. Cengage Learning, 2009.

[100] D. A. Forsyth, O. Arikan, and L. Ikemoto. *Computational Studies of Human Motion: Tracking and Motion Synthesis*. Now Publishers Inc, 2006.

[101] G. B. Frank E. Grubbs. Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, 1972.

[102] A. Gaidon, Z. Harchaoui, and C. Schmid. Recognizing activities with cluster-trees of tracklets. In *BMVC 2012-British Machine Vision Conference*, pages 30–1. BMVA Press, 2012.

[103] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[104] T. Gautama and M. M. Van Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *Neural Networks, IEEE Transactions on*, 13(5):1127–1136, 2002.

[105] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[106] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2913–2920. IEEE, 2009.

[107] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):1003–1016, 2012.

[108] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis. Representation and recognition of events in surveillance video using petri nets. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 112–112. IEEE, 2004.

[109] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 925–931. IEEE, 2009.

[110] S. Gong and T. Xiang. *Visual analysis of behaviour: from pixels to semantics*. Springer Science & Business Media, 2011.

[111] J. Gonzàlez, J. Varona, F. X. Roca, and J. J. Villanueva. aspaces: Action spaces for recognition and synthesis of human actions. In *Articulated Motion and Deformable Objects*, pages 189–200. Springer, 2002.

[112] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *Proc. BMVC*, pages 6.1–6.10, 2006. doi:10.5244/C.20.6.

[113] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[114] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2141–2148. IEEE, 2010.

[115] S. J. F. Guimarães, M. Couprie, A. de Albuquerque Araújo, and N. J. Leite. Video segmentation based on 2d image analysis. *Pattern Recognition Letters*, 24(7):947–957, 2003.

[116] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1148–1153. IEEE, 2006.

[117] H. Gunes, M. Piccardi, and T. Jan. Face and body gesture recognition for a vision-based multimodal analyzer. In *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing*, VIP '05, pages 19–28, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[118] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, volume 2, pages 325–329. IEEE, 1994.

[119] I. Guyon and A. Elisseeff. An introduction to feature extraction. In I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 1–25. Springer Berlin Heidelberg, 2006.

[120] E. T. Hall. *The silent language.* Anchor, 1973.

[121] I. Haritaoglu, D. Harwood, and L. Davis. W4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809–830, Aug 2000.

[122] J. Harrigan, R. Rosenthal, and K. Scherer. *New Handbook of Methods in Nonverbal Behavior Research*. Series in affective science. OUP Oxford, 2008.

[123] R. Hartley et al. In defense of the eight-point algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(6):580–593, 1997.

[124] T. Heibeck and A. Pentland. *Honest Signals: How They Shape Our World*. MIT Press, 2010.

[125] F. C. Heilbron, A. Thabet, J. C. Niebles, and B. Ghanem. Camera motion and surrounding scene appearance as context for action recognition. In *Computer Vision–ACCV 2014*, pages 583–597. Springer, 2014.

[126] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[127] A. Heloir and S. Gibet. A qualitative and quantitative characterisation of style in sign language gestures. In M. Sales Dias, S. Gibet, M. Wanderley, and R. Bastos, editors, *Gesture-Based Human-Computer Interaction and Simulation*, volume 5085 of *Lecture Notes in Computer Science*, pages 122–133. Springer Berlin Heidelberg, 2009.

[128] M. Hoai and A. Zisserman. Improving human action recognition using score distribution and ranking. In *Computer Vision–ACCV 2014*, pages 3–20. Springer, 2015.

[129] B. K. P. Horn. *Robot vision*. Cambridge: MIT Press, 1986.

[130] Y.-L. Hou and G. K. Pang. People counting and human detection in a challenging situation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41(1):24–33, 2011.

[131] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22Nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 482–500, Berlin, Heidelberg, 2010. Springer-Verlag.

[132] M. Hu, S. Ali, and M. Shah. Detecting global motion patterns in complex videos. In *ICPR*, pages 1–5. IEEE, 2008.

[133] M. Hu, S. Ali, and M. Shah. Learning motion patterns in crowded scenes using motion flow field. In *in Proceedings of the 19th International Conference on Pattern Recognition (ICPR?08*. Citeseer, IEEE, 2008.

[134] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3):334–352, 2004.

[135] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. J. Russell, and J. Weber. Automatic symbolic traffic scene analysis using belief networks. In *AAAI*, volume 94, pages 966–972, 1994.

[136] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[137] M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.*, 52(12):5186–5201, Aug. 2008.

[138] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 231–238, New York, NY, USA, 2011. ACM.

[139] B.-W. Hwang, S. Kim, and S.-W. Lee. 2d and 3d full-body gesture database for analyzing daily human gestures. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, editors, *Advances in Intelligent Computing*, volume 3644 of *Lecture Notes in Computer Science*, pages 611–620. Springer Berlin Heidelberg, 2005.

[140] Y. S. Isarun Chamveha, Yusuke Sugano and A. Sugimoto. Social group discovery from surveillance videos: A data-driven approach with attention-based cues. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.

[141] Y. Ivanov, A. F. Bobick, et al. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, 2000.

[142] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *Computer Vision–ECCV 2012*, pages 430–444. Springer, 2012.

[143] A. Jain, S. Chatterjee, and R. Vidal. Coarse-to-fine semantic video segmentation using supervoxel trees. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1865–1872. IEEE, 2013.

[144] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2555–2562. IEEE, 2013.

[145] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012.

[146] O. C. Jenkins and M. J. Mataric. Automated modularization of human motion into actions and behaviors. 2002.

[147] A. R. Jensenius. Using motiongrams in the study of musical gestures. In *Proceedings of the International Computer Music Conference*, pages 499–502, New Orleans, LA, 2006. Tulane University.

[148] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *Computer Vision–ECCV 2012*, pages 425–438. Springer, 2012.

[149] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 425–438. Springer Berlin Heidelberg, 2012.

[150] Y.-G. Jiang, X. Zeng, G. Ye, D. Ellis, S.-F. Chang, S. Bhattacharya, and M. Shah. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *TRECVID*, volume 20, pages 21–32, 2010.

[151] B. Jin, W. Hu, and H. Wang. Human interaction recognition based on transformation of spatial semantics. *IEEE Signal Process. Lett.*, 19(3):139–142, 2012.

[152] B. Jobard, G. Erlebacher, and M. Y. Hussaini. Lagrangian-eulerian advection for unsteady flow visualization. In *Proceedings of the Conference on Visualization '01*, VIS '01, pages 53–60, Washington, DC, USA, 2001. IEEE Computer Society.

[153] B. Jobard and W. Lefer. Creating evenly-spaced streamlines of arbitrary density. In *Eurographics Workshop*, pages 43–56. Springer Verlag, 1997.

[154] G. Johansson. Visual perception of biological motion and a model for its analysis. *Attention, Perception, & Psychophysics*, 14(2):201–211, 1973.

[155] K. Johnson and M. Shiffrar. *People watching: Social, perceptual, and neurophysiological studies of body perception*. Oxford University Press, 2013.

[156] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and vision computing*, 14(8):609–615, 1996.

[157] S.-B. Jun, K. Yoon, and H.-Y. Lee. Dissolve transition detection algorithm using spatio-temporal distribution of mpeg macro-block types (poster session). In *Proceedings of the eighth ACM international conference on Multimedia*, pages 391–394. ACM, 2000.

[158] Z. Kalal, J. Matas, and K. Mikolajczyk. Tracking learning detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[159] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *Pattern recognition (ICPR), 2010 20th international conference on*, pages 2756–2759. IEEE, 2010.

[160] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[161] G. Karypis. *Cluto: A Clustering Toolkit*. University of Minnesota, Department of Computer Science, November 2003.

[162] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013.

[163] S. Khalid and A. Naftel. Classifying spatiotemporal object trajectories using unsupervised learning of basis function coefficients. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 45–52. ACM, 2005.

[164] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, pages 2921–2928. IEEE, 2009.

[165] A. Kimber. Tests for many outliers in an exponential sample. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):263–271, 1982.

[166] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[167] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[168] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33, 2013.

[169] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(4):1027–1038, 2011.

[170] F. Klügl and G. Rindsfüser. Large-scale agent-based pedestrian simulation. In *Multiagent System Technologies*, pages 145–156. Springer, 2007.

[171] M. Knapp, J. Hall, and T. Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[172] B. Knyazev. Human nonverbal behavior multi-sourced ontological annotation. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, VIGTA '13, pages 2:1–2:8, New York, NY, USA, 2013. ACM.

[173] T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE*, pages 1–8. IEEE, 2008.

[174] Y. Kobayashi. The emotion sign: Human motion analysis classifying specific emotion. *JCP*, 3(9):20–28, 2008.

[175] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1446–1453. IEEE, 2009.

[176] V. Krüger, D. Kragic, A. Ude, and C. Geib. The meaning of action: a review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.

[177] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[178] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[179] N. Kumar and A. Vaish. Dominant flow based attribute grouping for indifferent movement detection in crowd. *International Journal of Computer Applications*, 88(18), 2014.

[180] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa. Automated detection of human for visual surveillance system. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 3, pages 865–869. IEEE, 1996.

[181] T. Lan, Y. Wang, G. Mori, and S. N. Robinovitch. Retrieving actions in group contexts. In *Trends and Topics in Computer Vision*, pages 181–194. Springer, 2012.

[182] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in neural information processing systems*, pages 1216–1224, 2010.

[183] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. *arXiv preprint arXiv:1411.6660*, 2014.

[184] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. *Double fusion for multimedia event detection*. Springer, 2012.

[185] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[186] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[187] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[188] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[189] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 120–127. IEEE, 2011.

[190] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[191] C. Li, Z. Han, Q. Ye, and J. Jiao. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neurocomputing*, 119:94–100, 2013.

[192] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[193] H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, 44(8):1614 – 1628, 2011.

[194] W. Liang, H. Weiming, and T. Tieniu. A survey of visual analysis of human motion. *Chinese Journal of Computers*, 25(3):225–237, 2002.

[195] R. W. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Electronic Imaging'99*, pages 290–301. International Society for Optics and Photonics, 1998.

[196] D. Lin, E. Grimson, and J. Fisher. Learning visual flows: A lie algebraic approach. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 747–754. IEEE, 2009.

[197] A. J. Lipton. *Local application of optic flow to analyse rigid versus non-rigid motion*. Carnegie Mellon University, The Robotics Institute, 1999.

[198] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pages 8–14. IEEE, 1998.

[199] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.

[200] D. Liu, H. Zhang, M. M. Polycarpou, C. Alippi, and H. He, editors. *Advances in Neural Networks - ISNN 2011 - 8th International Symposium on Neural Networks, ISNN 2011, Guilin, China, May 29-June 1, 2011, Proceedings, Part II*, volume 6676 of *Lecture Notes in Computer Science*. Springer, 2011.

[201] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE, 2009.

[202] J. Liu, S. McCloskey, and Y. Liu. Local expert forest of score fusion for video event classification. In *Computer Vision–ECCV 2012*, pages 397–410. Springer, 2012.

[203] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[204] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[205] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.

[206] S. Ma and W. Wang. Effective camera motion analysis approach. In *Networking, Sensing and Control (ICNSC), 2010 International Conference on*, pages 111–116. IEEE, 2010.

[207] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 469–478. ACM, 2012.

[208] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(3):397–408, 2005.

[209] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[210] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[211] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217, 1980.

[212] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.

[213] J. Mas and G. Fernandez. Video shot boundary detection based on color histogram. *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST*, 2003.

[214] J. Matas, C. Galambos, and J. Kittler. Robust detection of lines using the progressive probabilistic hough transform. *Computer Vision and Image Understanding*, 78(1):119–137, 2000.

[215] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 514–521. IEEE, 2009.

[216] T. Matsuo and S. Nakajima. Nikon multimedia event detection system. In *TRECVID*, 2010.

[217] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking group of people. *Computer Vision and Image Understanding*, 80:42–56, 2000.

[218] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[219] C. McPhail and R. T. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods and Research*, 10(3):347–375, 1982.

[220] A. Mebarki, P. Alliez, and O. Devillers. Farthest point seeding for efficient placement of streamlines. *Visualization Conference, IEEE*, 0:61, 2005.

[221] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder. Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16518–16523, 2005.

[222] R. Mehran, B. E. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (3)*, volume 6313 of *Lecture Notes in Computer Science*, pages 439–452. Springer, 2010.

[223] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009.

[224] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[225] D. Metaxas and S. Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6):421–433, 2013.

[226] Y. Mileva, A. Bruhn, and J. Weickert. Illumination-robust variational optical flow with photometric invariants. In *Proceedings of the 29th DAGM conference on Pattern recognition*, pages 152–162, Berlin, Heidelberg, 2007. Springer-Verlag.

[227] T. B. Moeslund and F. Bajers. Computer vision-based human motion capture - a survey, 1999.

[228] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268, 2001.

[229] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126, 2006.

[230] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(4):349–361, 2001.

[231] S. Molina-Giraldo, J. Carvajal-González, A. Álvarez-Meza, and G. Castellanos-Domínguez. Video segmentation framework based on multi-kernel representations and feature relevance analysis for object classification. In *Pattern Recognition Applications and Methods*, pages 273–283. Springer, 2015.

[232] S. Molina-Giraldo, H. Insuasti-Ceballos, C. Arroyave, J. Montoya, J. Lopez-Villa, A. Alvarez-Meza, and G. Castellanos-Dominguez. People detection in video streams using background subtraction and spatial-based scene modeling.

[233] B. Morris and M. Trivedi. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 312–319. IEEE, 2009.

[234] B. T. Morris and M. M. Trivedi. Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis. In *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*, pages 154–161. IEEE, 2008.

[235] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8):1114–1127, 2008.

[236] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409, 1981.

[237] J. M. Nadal, P. Monreal, and S. Perera. Emotion and linguistic diversity. *Procedia - Social and Behavioral Sciences*, 82(0):614 – 620, 2013.

[238] H.-H. Nagel. From image sequences towards conceptual descriptions. *Image and vision computing*, 6(2):59–74, 1988.

[239] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003.

[240] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, U. Park, R. Prasad, and P. Natarajan. Multi-channel shape-flow kernel descriptors for robust video event detection and retrieval. In *Computer Vision–ECCV 2012*, pages 301–314. Springer, 2012.

[241] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1298–1305. IEEE, 2012.

[242] A. Natsev, J. R. Smith, M. Hill, G. Hua, B. Huang, M. Merler, L. Xie, H. Ouyang, and M. Zhou. Ibm research trecvid-2010 video copy detection and multimedia event detection system. In *Proceedings of NIST TRECVID, Workshop*. Citeseer, 2010.

[243] J. Ng and S. Gong. Learning pixel-wise signal energy for understanding semantics. In *In Proc. BMVC*, pages 695–704. Press, 2001.

[244] B. Ni, P. Moulin, X. Yang, and S. Yan. Motion part regularization: Improving action recognition via trajectory selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3698–3706, 2015.

[245] B. Ni, S. Yan, and A. Kassim. Recognizing human group activities with localized causalities. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1470–1477. IEEE, 2009.

[246] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Computer Vision–ECCV 2010*, pages 392–405. Springer, 2010.

[247] M. Nieto, C. Cuevas, L. Salgado, and N. N. García. Line segment detection using weighted mean shift procedures on a 2d slice sampling strategy. *Pattern Anal. Appl.*, 14(2):149–163, 2011.

[248] M. Nieto and L. Salgado. Non-linear optimization for robust estimation of vanishing points. In *ICIP*, pages 1885–1888. IEEE, 2010.

[249] A. R. N. Nilchi and M. R. . An efficient algorithm for motion detection based facial expression recognition using optical flow. *International Journal of Engineering and Applied Sciences*, 14:318–323, 2006.

[250] T. Nir, A. M. Bruckstein, and R. Kimmel. Over-parameterized variational optical flow. *Int. J. Comput. Vision*, 76(2):205–216, Feb. 2008.

[251] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*. Springer, 2006.

[252] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1187–1200, June 2014.

[253] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of visual communication and image representation*, 6(4):348–365, 1995.

[254] S. Oh and A. Hoogs. Unsupervised learning of activities in video using scene context. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3579–3582. IEEE, 2010.

[255] A. Oikonomopoulos and M. Pantic. Human activity recognition using hierarchically-mined feature constellations. In *Advances in Visual Computing*, pages 150–159. Springer, 2013.

[256] A. Oikonomopoulos, I. Patras, M. Pantic, and N. Paragios. Trajectory-based representation of human actions. In *Proceedings of the ICMI 2006 and IJCAI 2007 International Conference on Artifical Intelligence for Human Computing*, ICMI'06/IJCAI'07, pages 133–154, Berlin, Heidelberg, 2007. Springer-Verlag.

[257] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[258] A. Oliva, A. Torralba, A. Guérin-Dugué, and J. Hérault. Global semantic classification of scenes using power spectrum templates. In *Proceedings of the 1999 international conference on Challenge of Image Retrieval*, pages 9–9. British Computer Society, 1999.

[259] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1817–1824. IEEE, 2013.

[260] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, volume 1, page 3, 2009.

[261] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[262] J. Owens and A. Hunter. Application of the self-organizing map to trajectory classification. In *Proceedings of the Third IEEE International Workshop on Visual Surveillance (VS'2000)*, VS '00, pages 77–, Washington, DC, USA, 2000. IEEE Computer Society.

[263] O. Ozturk, T. Yamasaki, and K. Aizawa. Detecting dominant motion flows in unstructured/structured crowd scenes. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pages 3533–3536, Washington, DC, USA, 2010. IEEE Computer Society.

[264] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang. Human computing and machine understanding of human behavior: a survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[265] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, pages 1370–1390, 2003.

[266] A. N. Papadopoulos. Trajectory retrieval with latent semantic analysis. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1089–1094. ACM, 2008.

[267] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *Int. J. Comput. Vision*, 67(2):141–158, Apr. 2006.

[268] A. J. Patti, M. I. Sezan, et al. A new motion-compensated reduced-order model kalman filter for space-varying restoration of progressive and interlaced video. *Image Processing, IEEE Transactions on*, 7(4):543–554, 1998.

[269] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[270] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 261–268. IEEE, 2009.

[271] X. Peng, Y. Qiao, Q. Peng, and X. Qi. Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition. In *British Machine Vision Conference (BMVC)*, 2013.

[272] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv preprint arXiv:1405.4506*, 2014.

[273] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *Computer Vision–ECCV 2014*, pages 581–595. Springer, 2014.

[274] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):107–119, 2000.

[275] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, volume 5, 2004.

[276] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 733–740. IEEE, 2012.

[277] E. M. Pereira, J. S. Cardoso, and R. Morla. A critical analysis about a motion-based approach to extract global trajectories. Nov 2013.

[278] E. M. Pereira, J. S. Cardoso, and R. Morla. Motion flow tracking in unconstrained videos for retail scenario. In J. a. Sanches, L. Micó, and J. S. Cardoso, editors, *Pattern Recognition and Image Analysis*, volume 7887 of *Lecture Notes in Computer Science*, pages 340–349. Springer Berlin Heidelberg, 2013.

[279] E. M. Pereira, J. S. Cardoso, and R. Morla. Long-range trajectories from global and local motion representations. Sep 2015.

[280] E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Context-based trajectory descriptor for human activity profiling. In *IProc. IEEE Int. Conf. Syst., Man, Cybern., San Diego, CA, USA*, pages 2385–2390. IEEE, IEEE, Oct 2014.

[281] E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Social signaling descriptor for group behavior analysis. In R. Paredes, X. M. Pardo, and J. S. Cardoso, editors, *Pattern Recognition and Image Analysis*, volume 9117 of *Lecture Notes in Computer Science*, pages 13–22, June 2015.

[282] E. M. Pereira, L. Ciobanu, and J. S. Cardoso. Cross-layer classification framework for automatic social behaviour analysis in surveillance scenario. *Neurocomputing*, 2016.

[283] E. M. Pereira, S. Molina-Giraldo, L. Ciobanu, D. Insuasti-Ceballos, A. Álvarez Meza, G. Castellanos-Dominguez, and J. S. Cardoso. Group discovery and feature relevance analysis for social behavior identification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2016.

[284] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[285] S. Pfeiffer, R. Lienhart, G. Kühne, and W. Effelsberg. The moca project. In *Informatik?98*, pages 329–338. Springer, 1998.

[286] S. Phan, T. D. Ngo, V. Lam, S. Tran, D.-D. Le, D. A. Duong, and S. Satoh. Multimedia event detection using segment-based approach for motion feature. *Journal of Signal Processing Systems*, 74(1):19–31, 2014.

[287] S. Piana, M. Mancini, A. Camurri, G. Varni, and G. Volpe. Automated analysis of non-verbal expressive gesture. In *Human Aspects in Ambient Intelligence*, pages 41–54. Springer, 2013.

[288] S. Piana, A. Staglianó, A. Camurri, and F. Odone. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. *IDGEI International Workshop*, 2013.

[289] J. Pinto da Costa, R. Sousa, and J. Cardoso. An all-at-once unimodal svm approach for ordinal classification. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 59–64, Dec 2010.

[290] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2537–2544. IEEE, 2014.

[291] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51 – B61, 2001.

[292] M. Pomplun, H. Ritter, and B. Velichkovsky. Disambiguating complex visual information: Towards communication of personal views of a scene. *PERCEPTION-LONDON-*, 25:931–948, 1996.

[293] R. Poppe. Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18, 2007.

[294] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[295] A. B. Poritz. Hidden markov models: A guided tour. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 7–13. IEEE, 1988.

[296] B. L. Price, B. S. Morse, and S. Cohen. Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 779–786. IEEE, 2009.

[297] V. Prisacariu and I. Reid. fasthog - a real-time gpu implementation of hog. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009.

[298] A. Psarrou, S. Gong, and M. Walter. Recognition of human gestures and behaviour based on motion trajectories. *Image and Vision Computing*, 20(5):349–358, 2002.

[299] G. Pusiol, F. Bremond, and M. Thonnat. Trajectory based activity discovery. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 270–277. IEEE, 2010.

[300] S. R. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *In IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[301] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proceedings of the 11th European Conference on Computer Vision: Part I*, ECCV'10, pages 577–590, Berlin, Heidelberg, 2010. Springer-Verlag.

[302] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[303] W.-y. Ren, G.-h. Li, J. Chen, and H.-z. Liang. Abnormal crowd behavior detection using behavior entropy model. In *Wavelet Analysis and Pattern Recognition (ICWAPR), 2012 International Conference on*, pages 212–221. IEEE, 2012.

[304] P. C. Ribeiro and J. Santos-Victor. Human activity recognition from video: modeling, feature selection and classification architecture. In *Proceedings of International Workshop on Human Activity Recognition and Modelling*, pages 61–78. Citeseer, 2005.

[305] V. P. Richmond, J. C. McCroskey, and S. K. Payne. *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

[306] L. Riek, A. S., and R. P. Affect decoding measures and human-computer interaction. *In Proceedings of Measuring Behaviour*, 2008.

[307] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[308] I. V. Rodrigues, E. M. Pereira, and L. F. Teixeira. Analysis of expressiveness of portuguese sign language speakers. In R. Paredes, X. M. Pardo, and J. S. Cardoso, editors, *Pattern Recognition and Image Analysis*, volume 9117 of *Lecture Notes in Computer Science*, pages 708–717, June 2015.

[309] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[310] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *In European Conference on Computer Vision*, pages 430–443, 2006.

[311] K. K. Roudposhti and J. Dias. Probabilistic human interaction understanding: Exploring relationship between human body motion and the environmental context. *Pattern Recognition Letters*, 34(7):820 – 830, 2013. Scene Understanding and Behaviour Analysis.

[312] H. A. Rowley and J. M. Rehg. Analyzing articulated motion using expectation-maximization. In *cvpr*, page 935. IEEE, 1997.

[313] M. Rubinstein, C. Liu, and W. T. Freeman. Towards longer long-range motion trajectories. pages 53.1–53.11, 2012.

[314] D. M. Russel and S. Gong. Minimum cuts of a time-varying background. *British Machine Vision Association*, pages 809–818, 2006.

[315] N. Russo. Connotation of seating arrangements. *Cornell Journal of Social Relations*, 2(1):37–44, 1967.

[316] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Computer vision and pattern recognition, 2006 ieee computer society conference on*, volume 2, pages 1709–1718. IEEE, 2006.

[317] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, pages 1593–1600. IEEE, 2009.

[318] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1472–1485, 2009.

[319] M. Sales Dias, S. Gibet, M. Wanderley, and R. Bastos. *Gesture-Based Human-Computer Interaction and Simulation, Proceedings of Gesture Workshop 2007*. Lecture Notes in Computer Science. Springer, Dec. 2009.

[320] P. Sand and S. J. Teller. Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 80(1):72–91, 2008.

[321] D. Schachner, P. Shaver, and M. Mikulincer. Patterns of nonverbal behavior and sensivity in the context of attachment relations. *Journal of Nonverbal Behavior*, 29(3):141–169, 2005.

[322] A. E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27(4):316–331, 1964.

[323] A. Scherp and V. Mezaris. Survey on modeling and indexing events in multimedia. *Multimedia Tools and Applications*, 70(1):7–23, 2014.

[324] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[325] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360. ACM, 2007.

[326] P. Scovanner and M. F. Tappen. Learning pedestrian dynamics from the real world. In *ICCV*, volume 9, pages 381–388, 2009.

[327] A. W. Senior, L. Brown, A. Hampapur, C.-F. Shu, Y. Zhai, R. S. Feris, Y.-L. Tian, S. Borger, and C. Carlson. Video analytics for retail. 2007.

[328] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3547–3551, Sept 2013.

[329] M. Shah. Understanding human behavior from motion imagery. *Machine Vision and Applications*, 14(4):210–214, 2003.

[330] M. Shah and R. Jain. Visual recognition of activities, gestures, facial expressions and speech: an introduction and a perspective. In *Motion-Based Recognition*, pages 1–14. Springer, 1997.

[331] L. Shao, S. Jones, and X. Li. Efficient search and localization of human actions in video databases. *IEEE Trans. Circuits Syst. Video Techn.*, 24(3):504–512, 2014.

[332] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 593 –600, jun 1994.

[333] M. H. Siddiqi, R. Ali, M. S. Rana, E.-K. Hong, E. S. Kim, and S. Lee. Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. *Sensors*, 14(4):6370–6392, 2014.

[334] A. F. Smeaton, P. Over, and A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[335] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330. ACM, 2006.

[336] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba. Evaluating multi-object tracking. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 36–36, June 2005.

[337] L. I. Smith. A tutorial on principal components analysis. Technical report, Cornell University, USA, February 26 2002.

[338] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.

[339] B. Solmaz, S. M. Assari, and M. Shah. Classifying web videos using a global video descriptor. *Machine vision and applications*, 24(7):1473–1485, 2013.

[340] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757, 2000.

[341] E. Stringa. Morphological change detection algorithms for surveillance applications. In *BMVC*, pages 1–10, 2000.

[342] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 15–22. IEEE, 2013.

[343] D. Sun, J. P. Lewis, and M. Black. Learning optical flow. In *In Proc. ECCV*, pages 83–97, 2008.

[344] D. Sun, S. Roth, and M. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439, 2010.

[345] J. Sun, Y. Mu, S. Yan, and L. F. Cheong. Activity recognition using dense long-duration trajectories. In *ICME*, pages 322–327. IEEE, 2010.

[346] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2004–2011. IEEE, 2009.

[347] X. Sun, D. Huang, Y. Wang, and J. Qin. Action recognition based on kinematic representation of video data. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1530–1534. IEEE, 2014.

[348] M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In *Computer Vision and Pattern Recognition*, pages 9–16, 2011.

[349] M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. [paper] using trajectory features to recognize human actions within crowd sequences of real surveillance video. *ITE Transactions on Media Technology and Applications*, 1(2):118–126, 2013.

[350] R. Thompson. A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):53–55, 1985.

[351] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2642–2649. IEEE, 2013.

[352] S. Tomkins. *Affect Imagery Consciousness: Volume II: The Negative Affects*. Springer Series. Springer Publishing Company, 1963.

[353] S. S. Tomkins. *Affect Imagery Consciousness: Volume I The Positive Affects*. New York Springer Publishing, 1962.

[354] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.

[355] M. Valera and S. A. Velastin. Intelligent distributed surveillance systems: a review. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 152, pages 192–204. IET, 2005.

[356] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[357] M. Van Erp, L. Vuurpijl, and L. Schomaker. An overview and comparison of voting methods for pattern recognition. In *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, pages 195–200. IEEE, 2002.

[358] V. Verma, D. Kao, and A. Pang. A flow-guided streamline seeding strategy. In *Proceedings of the conference on Visualization '00*, VIS '00, pages 163–170, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.

[359] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Computer Vision–ECCV 2012*, pages 84–97. Springer, 2012.

[360] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[361] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris. Matching mixtures of trajectories for human action recognition. *Computer Vision and Image Understanding*, 19:27–40, 2014.

[362] H. G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998.

[363] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[364] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[365] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, June 2011.

[366] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[367] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, pages 124–1. BMVA Press, 2009.

[368] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. *Computer Vision-ECCV 2004*, pages 238–249, 2004.

[369] J. J. Wang and S. Singh. Video analysis of human dynamics: a survey. *Real-time imaging*, 9(5):321–346, 2003.

[370] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern recognition*, 36(3):585–601, 2003.

[371] T. Wang, S. Wang, and X. Ding. Detecting human action as the spatio-temporal tube of maximum mutual information. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(2):277–290, 2014.

[372] X. Wang, K. T. Ma, G.-W. Ng, and W. E. Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *Int. J. Comput. Vision*, 95(3):287–312, Dec. 2011.

[373] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *Computer Vision–ECCV 2006*, pages 110–123. Springer, 2006.

[374] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *Computer Vision–ECCV 2006*, pages 110–123. Springer, 2006.

[375] T. Watanabe, S. Ito, and K. Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. In *Advances in Image and Video Technology*, pages 37–47. Springer, 2009.

[376] T. Weinkauf and H. Theisel. Streak lines as tangent curves of a derived vector field. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1225–1234, Nov. 2010.

[377] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[378] A. Whitehead, P. Bose, and R. Laganiere. Feature based cut detection with automatic threshold selection. In *Image and Video Retrieval*, pages 410–418. Springer, 2004.

[379] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda. A suspicious behaviour detection using a context space model for smart surveillance systems. *Computer Vision and Image Understanding*, 116(2):194 – 209, 2012.

[380] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[381] J. Wu and J. M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *2009 IEEE 12th International Conference on Computer Vision*, pages 630–637. IEEE, 2009.

[382] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1489–1501, 2011.

[383] K. Wu, Z. Liu, S. Zhang, and R. J. Moorhead II. Topology-aware evenly spaced streamline placement. *IEEE Transactions on Visualization and Computer Graphics*, 16(5):791–801, Sept. 2010.

[384] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2054–2060. IEEE, 2010.

[385] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1419–1426. IEEE, 2011.

[386] Y. Wu and T. S. Huang. Human hand modeling, analysis and animation in the context of hci. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 3, pages 6–10. IEEE, 1999.

[387] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[388] T. Xiang and S. Gong. Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67:2006, 2006.

[389] T. Xiang, S. Gong, and D. Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. *In British Machine Vision Conference*, pages 233–242, 2002.

[390] J. Xiao, H. Cheng, H. Sawhney, C. Rao, M. Isnardi, and S. Corporation. Bilateral filtering-based optical flow estimation with occlusion detection. In *In ECCV, volume I*, pages 211–224, 2006.

[391] W. Xiong and J. C.-M. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, 1998.

[392] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1202–1209. IEEE, 2012.

[393] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1985–1997, 2008.

[394] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, and A. G. Hauptmann. Event detection using multi-level relevance labels and multiple features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 97–104. IEEE, 2014.

[395] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. *arXiv preprint arXiv:1411.4006*, 2014.

[396] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. In *Computer Vision, 1998. Sixth International Conference on*, pages 120–127. IEEE, 1998.

[397] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(6):636–642, 1996.

[398] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113, 2008.

[399] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *In ECCV*, pages 94–106, 2006.

[400] Y. Yanagisawa and T. Satoh. Clustering multidimensional trajectories based on shape and velocity. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, ICDEW '06, pages 12–, Washington, DC, USA, 2006. IEEE Computer Society.

[401] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2104–2111. IEEE, 2013.

[402] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li. Human action recognition by learning bases of action attributes and parts. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, editors, *ICCV*, pages 1331–1338. IEEE, 2011.

[403] F. Yuan, V. Prinet, and J. Yuan. Middle-level representation for human activities recognition: the role of spatio-temporal relationships. In *Trends and Topics in Computer Vision*, pages 168–180. Springer, 2012.

[404] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1728–1743, 2011.

[405] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(2):168–186, 2007.

[406] J. yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.

[407] S. Zaidenberg, B. Boulay, C. Garate, D. P. Chau, E. Corvee, and F. Bremond. Group interaction and group tracking for video-surveillance in underground railway stations. In *International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011)*, page 10, Sophia Antipolis, France, Sept. 2011.

[408] L. Zelnik-manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.

[409] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space Speaks - Towards Socially and Personality Aware Visual Surveillance. In *ACM MM'10 Workshop, Multimodal Pervasive Video Analysis Workshop (MPVA), Firenze, Italy*, 2010.

[410] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[411] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, 2008.

[412] Y. Zhang, W. Ge, M.-C. Chang, and X. Liu. Group context learning for event recognition. In *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, pages 249–255. IEEE, 2012.

[413] X. Zhao, D. Gong, and G. Medioni. Tracking using motion patterns for very crowded scenes. In *Proceedings of the 12th European conference on Computer Vision - Volume Part II*, ECCV'12, pages 315–328, Berlin, Heidelberg, 2012. Springer-Verlag.

[414] B. Zhou, X. Tang, and X. Wang. Coherent filtering: Detecting coherent motions from crowd clutters. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *ECCV (2)*, volume 7573 of *Lecture Notes in Computer Science*, pages 857–871. Springer, 2012.

[415] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *CVPR*, pages 2871–2878. IEEE, 2012.

[416] H. Zimmer, A. Bruhn, J. Weickert, L. Valgaerts, A. Salgado, B. Rosenhahn, and H.-P. Seidel. Complementary optic flow. In *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, EMMCVPR '09, pages 207–220, Berlin, Heidelberg, 2009. Springer-Verlag.

[417] C. L. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1308–1315. IEEE, 2005.

Appendix A

Appendix

Here in the Appendix are attached some auxiliary information about the methodologies described in Chapter 6.

As seguintes afirmações por favor, indique em que medida concorda com cada uma das seguintes afirmações.

No espaço anterior a cada afirmação, escreva o número (1, 2, 3, 4, 5, 6, 7, 8, ou 9) que indica em que medida as seguintes afirmações descrevem a sua opinião. Quando responder, por favor procure utilizar todos os números desta escala (1 a 9). Não há respostas corretas ou erradas a estas questões.

Discordo totalmente 1	2	Discordo moderadamente 3	4	Neutro 5	6	Concordo moderadamente 7	8	Concordo totalmente 9

_____ Considero importante a formação profissional dos surdos

_____A comunidade ouvinte não conhece a comunidade surda.

_____A comunidade ouvinte não se preocupa com a comunidade surda.

_____Existe igualdade de oportunidades educativas/profissionais para surdos e ouvintes.

_____A comunidade ouvinte descrimina a comunidade surda.

_____Existe preconceito em relação à pessoa surda.

_____A Língua Gestual Portuguesa deveria ser lecionada para os alunos em geral.

_____Os surdos e os ouvintes têm os mesmos direitos.

_____Considero que os surdos devem frequentar as mesmas escolas que ouvintes.

INSTRUÇÕES: Por favor, indique os seguintes dados demográficos.

1. Nome _____.
2. Idade _____.
3. Sexo:
   _____ Masculino
   _____ Feminino
4. Qual a sua naturalidade?
   _____ Portuguesa
   _____Outra
5. Qual o grau de escolaridade que concluiu?

_____ Sem escolaridade
_____ Ensino primário
_____ Ensino básico
_____ Ensino secundário
_____ Ensino superior
_____ Ensino profissional
_____Outro. Qual?_____.

6. Qual a sua situação atual perante o emprego?

_____Empregado

_____Desempregado

_____Reformado.

_____Estudante.

7. Qual a sua profissão?_____.

8. Conhece a Língua Gestual Portuguesa?

_____Não.

_____Sim.

Se sim, há quanto tempo? _____.

9. Qual a sua formação ao nível da Língua Gestual Portuguesa (LGP)?

____docência de língua gestual portuguesa.

____interpretação de língua gestual portuguesa.

____cursos básicos de língua gestual portuguesa (por ex. terapia da fala).

____não tem.

10. Tem conhecimento da comunidade surda ou de associações representativas das pessoas surdas?

_____Não.

_____Sim, mas poucas.

_____Sim, muitas.

11. Tem amigos surdos?

_____Não.

_____Sim, mas poucos.

_____Sim, muitos.

12. Nasceu surdo?

_____Sim.

_____Não.

Se não, com que idade ficou surdo(a)?_____. E qual a razão?_____.

13. Qual o seu tipo de surdez?

_____.

14. Os seus pais são ouvintes ou surdos?

_____Ouvintes

_____Surdos

_____Ouvinte/Surdo.

15. Tem mais familiares com surdez?

_____Não.

_____Sim.

16. Usa prótese auditiva?

_____Sim.

_____Não, nunca.

_____Não, mas já usei.

17. Usa implante?

_____Não.

_____Sim.

Agradecemos a sua participação.

Figure A.1: Questionnaire used for deaf individuals containing demographic and opinion questions. Portuguese version of the questionnaire.

As seguintes afirmações por favor, indique em que medida concorda com cada uma das seguintes afirmações.

No espaço anterior a cada afirmação, escreva o número (1, 2, 3, 4, 5, 6, 7, 8, ou 9) que indica em que medida as seguintes afirmações descrevem a sua opinião. Quando responder, por favor procure utilizar todos os números desta escala (1 a 9). Não há respostas corretas ou erradas a estas questões.

Discordo totalmente 1	2	Discordo moderadamente 3	4	Neutro 5	6	Concordo moderadamente 7	8	Concordo totalmente 9

_____ Considero importante a formação profissional dos surdos

_____A comunidade ouvinte não conhece a comunidade surda.

_____A comunidade ouvinte não se preocupa com a comunidade surda.

_____Existe igualdade de oportunidades educativas/profissionais para surdos e ouvintes.

_____A comunidade ouvinte descrimina a comunidade surda.

_____Existe preconceito em relação à pessoa surda.

_____A Língua Gestual Portuguesa deveria ser lecionada para os alunos em geral.

_____Os surdos e os ouvintes têm os mesmos direitos.

_____Considero que os surdos devem frequentar as mesmas escolas que ouvintes.

INSTRUÇÕES: Por favor, indique os seguintes dados demográficos.

1. Nome _____.
2. Idade _____.
3. Sexo:
   _____ Masculino
   _____ Feminino
4. Qual a sua naturalidade?
   _____ Portuguesa
   _____Outra

5. Qual o grau de escolaridade que concluiu?

_____ Sem escolaridade
_____ Ensino primário
_____ Ensino básico
_____ Ensino secundário
_____ Ensino superior
_____ Ensino profissional
_____Outro. Qual?_____.

6. Qual a sua situação atual perante o emprego?

_____Empregado

_____Desempregado

_____Reformado.

_____Estudante.

7. Qual a sua profissão?_____.

8. Conhece a Língua Gestual Portuguesa?

_____Não.

_____Sim.

Se sim, há quanto tempo? _____.

9. Qual a sua formação ao nível da Língua Gestual Portuguesa (LGP)?

____docência de língua gestual portuguesa.

____interpretação de língua gestual portuguesa.

____cursos básicos de língua gestual portuguesa (por ex. terapia da fala).

____não tem.

10. Tem conhecimento da comunidade surda ou de associações representativas das pessoas surdas?

_____Não.

_____Sim, mas poucas.

_____Sim, muitas.

11. Tem amigos surdos?

_____Não.

_____Sim, mas poucos.

_____Sim, muitos.

Agradecemos a sua participação.

Figure A.2: Questionnaire used for hearing individuals containing demographic and opinion questions. Portuguese version of the questionnaire.